Alma Mater Studiorum · Università di Bologna

DOTTORATO DI RICERCA IN
COMPUTER SCIENCE AND ENGINEERING
Ciclo XXXVI

Settore Concorsuale: 09/H1 - Sistemi di Elaborazione delle Informazioni
Settore Scientifico Disciplinare: INF/01 - Informatica

Mind the Gaps: Cognitive-Inspired AI for High-Level Visual Sensemaking

Towards Abstract Concept Image Classification

Presentata da:
Delfina Sol Martinez Pandiani

Coordinatore Dottorato:
Ilaria Bartolini

Supervisore:
Valentina Presutti

Co-Supervisore:
Sofia Pescarin

Esame Finale Anno 2024

Gracias a la vida
que me ha dado tanto
me dio dos luceros, que cuando los abro
perfecto distingo, lo negro del blanco
y en el alto cielo su fondo estrellado
y en las multitudes, la muter que yo amo.
Gracias a la vida
que me ha dado tanto
me ha dado el oído que en todo su ancho
graban noche y días, grillos y canarios
martillos, turbinas, ladridos, chubascos
y la voz tan tierna de mi bien amado.
Gracias a la vida
que me ha dado tanto
me ha dado el sonido y el abecedario
con él las palabras que pienso y declaro
madre amigue hermane, y luz alumbrando
la ruta del alma del que estoy amando.
Gracias a la vida
que me ha dado tanto
me ha dado la marcha de mis pies cansados
con ellos anduve ciudades y charcos
playas y desiertos, montañas y llanos
y la casa tuya, tu calle y tu patio.
Gracias a la vida
que me ha dado tanto
me dio el corazón que agita su marco
cuando miro el fruto del cerebro humano
cuando miro el bueno tan lejos del malo
cuando miro el fondo de tus ojos claros.
Gracias a la vida
que me ha dado tanto
me ha dado la risa, y me ha dado el llanto
así yo distingo dichas de quebrantos
los dos materiales que forman mi canto
y el canto de ustedes que es el mismo canto
y el canto de todos que es mi propio canto.
-Mercedes Sosa

Abstract

The abundance of visual data and the push for robust AI are driving the need for automated visual sensemaking. Computer Vision (CV) faces growing demand for models that can discern not only what images "represent", but also what they "evoke." This is a demand for tools mimicking human perception at a high semantic level, categorizing images based on abstract concepts like *freedom*, *danger*, or *safety*. However, automating this process is challenging due to entropy, scarcity, subjectivity, and ethical considerations. These challenges not only impact performance, but also underscore the critical need for interpretability, given existing semantic and cultural gaps between raw pixel data and high-level visual semantics.

This dissertation aims to bridge these gaps by focusing on abstract concept-based (AC) image classification, guided by three technical principles: situated grounding, performance enhancement, and interpretability. We introduce ART-stract, a novel dataset of cultural images annotated with ACs, serving as the foundation for a series of experiments across four key domains: assessing the effectiveness of the end-to-end DL paradigm, exploring cognitive-inspired semantic intermediaries, incorporating cultural and commonsense aspects, and neuro-symbolic integration of sensory-perceptual data with cognitive-based knowledge.

Our results demonstrate that integrating CV approaches with semantic technologies yields methods that surpass the current state of the art in AC image classification, outperforming the end-to-end deep vision paradigm both in regards to performance and to explainability. The results emphasize the role semantic technologies can play in developing both effective and interpretable systems, through the capturing, situating, and reasoning over knowledge related to visual data.

Furthermore, this dissertation explores the complex interplay between technical and socio-technical factors, emphasizing the importance of context in automatically interpreting visual data and cultural connotations. By merging technical expertise with an understanding of human and societal aspects, we advocate for responsible labeling and training practices in visual media. These insights and techniques not only advance efforts in CV and explainable artificial intelligence, but also propel us toward an era of AI development that harmonizes technical prowess with deep awareness of its human and societal implications.

Contents

List of Abbreviations

AC	Abstract Concept
AI	Artificial Intelligence
AKG	ARTstract Knowledge Graph
AM	Activation Maximization
CAM	Class Activation Mapping
CH	Cultural Heritage
CNN	Convolutional Neural Network
CV	Computer Vision
FV	Feature Visualization
KB	Knowledge Base
KG	Knowledge Graph
KGE	Knowledge Graph Embedding
DL	Deep Learning
ML	(Classical) Machine Learning
PS	Perceptual Semantics
SD-AM	Stable Diffusion Activation Maximization
SPK	Situated Perceptual Knowledge
VG	Visual Genome
ViT	Visual Transformer
VSKG	Visual Sense Knowledge Graph
VSO	Visual Sense Ontology
WAT	Words As Social Tools
XAI	eXplainable Artificial Intelligence

Part I

Background

Chapter I.1

Introduction

I.1.1 The Era of Algorithmic Curation

I.1.1.1 Curating by Abstracting

The modern mass media landscape, including vast digitalization efforts by Cultural Heritage (CH) institutions [47], is saturated with images. In this age of visual abundance, the sheer volume of diverse imagery leads to users' information overload [181]. Consequently, there is a growing need for society to complement traditional reading habits with a more visual perspective [334]. Amidst this visually overwhelming environment, *hyperpop* and glitch aesthetics (see Figure I.1.1) have emerged as noteworthy responses by the tech-savvy Generation Z, reflecting the quest for coherence in a world saturated with images. These aesthetics align with a culture characterized by rapid information flow and the pursuit of meaning within the visual chaos [399]. As suggested by Vassar, the hyperpop phenomenon is an attempt suited to the psyche of the modern "six-hours-of-screen-time-a-day individual," to strive for some semblance of meaning amidst the disorder; an attempt to coalesce a multitude of disparate meanings into some harmonious whole [366].

This is a transition from the information age to an era demanding selection and arrangement. In this shift, the ability to effectively sift through vast amounts of data is becoming the true source of intelligence and influence [300, 84, 181]. This idea was already at the core of Borges' premonitory short story "Funes the Memorious" (1954) [50], about Professor Funes, a man who possesses an extraordinary memory capable of recalling even the minutest details of his experiences. However, despite his exceptional memory, Borges describes Funes as "not very capable of thought" [50]. Borges explains this apparent contradiction by stating:

> *To think is to forget a difference, to generalize, to abstract.*

(a) Mikey Joyce 2020[a]

(b) Claire Barrow 2020[a]

[a] https://www.instagram.com/mikeyjoyce/ Access date: December 2023.

[a] https://www.instagram.com/claire_barrow/ Access date: December 2023.

Figure I.1.1: Hyperpop collapses disparate visual meanings into semblances of accord, symptomatic of an era characterized by visual overabundance.

Professor Funes' cognitive challenge, his inability to forget, impedes him from generalizing over interconnected memories. Illuminating these profound connections between memory, abstraction, and curation, Borges' story anticipated modern neuroscientific and philosophical research [207, 284, 325, 285], and now resonates with contemporary Artificial Intelligence (AI) research–currently dominated by Deep Learning (DL)–as it aims to achieve robust AI, as defined by Marcus [237]:

> *Intelligence that [...] can be counted on to apply what it knows to a wide range of problems in a systematic and reliable way, synthesizing knowledge from a variety of sources such that it can reason flexibly and dynamically about the world, transferring what it learns in one context to another, in the way that we would expect of an ordinary adult.*

This comprehensive vision of "robust" intelligence pursued by contemporary AI research signifies a departure from mere data accumulation toward the sophisticated synthesis and application of knowledge. However, similar to Funes, modern AI still lacks the abstraction abilities typically found in an "ordinary adult." In other words, despite their extensive data storage capabilities, neither Funes nor the DL paradigm have mastered the art of curation, which demands high-level reasoning and abstraction.

I.1.1.2 Automating High-Level Visual Sensemaking

Amidst visual hyperabundance and the rising popularity of AI, there is a heightened demand for automated curation, classification, and efficient navigation of this extensive visual terrain. This demand aligns with the evolving field of Computer Vision (CV). Initially developed with the goal of achieving "complete image understanding" [168], the field of CV treats images as data [157] and generally facilitates bottom-up access to extensive image repositories [18]. As such, it allows the simultaneous analysis of vast amounts of images, reducing the need for painstaking individual examination. As a field, CV encompasses the construction of physical models of scenes, the understanding of how light interacts with scenes, and the generation of low-, intermediate-, and high-level descriptions of scene content [349]. It also seeks high-level interpretation from a wide array of visual data [55] and aims to replicate the human capability of not only recovering image structure but also of identifying what an image "represents."

Over the past decade, the field of CV has undergone a profound transformation driven by the advent of DL, specifically powered by Convolutional Neural Networks (CNNs). DL techniques have harnessed the capabilities of extensive data and powerful computing resources to tackle once-considered insurmountable challenges, pushing the boundaries of what is achievable [275]. A compelling illustration of this paradigm shift can be seen in image classification, notably catalyzed by the breakthrough of Krizhevsky, Sutskever, and Hinton in 2012 [211]. Since then, DL has consistently outperformed traditional methods in this domain [275]. CNNs and other DL-based computer vision methods have now become indispensable tools, enabling the classification and categorization of extensive image datasets, even encompassing cultural imagery like advertisements [391] and artworks [71].

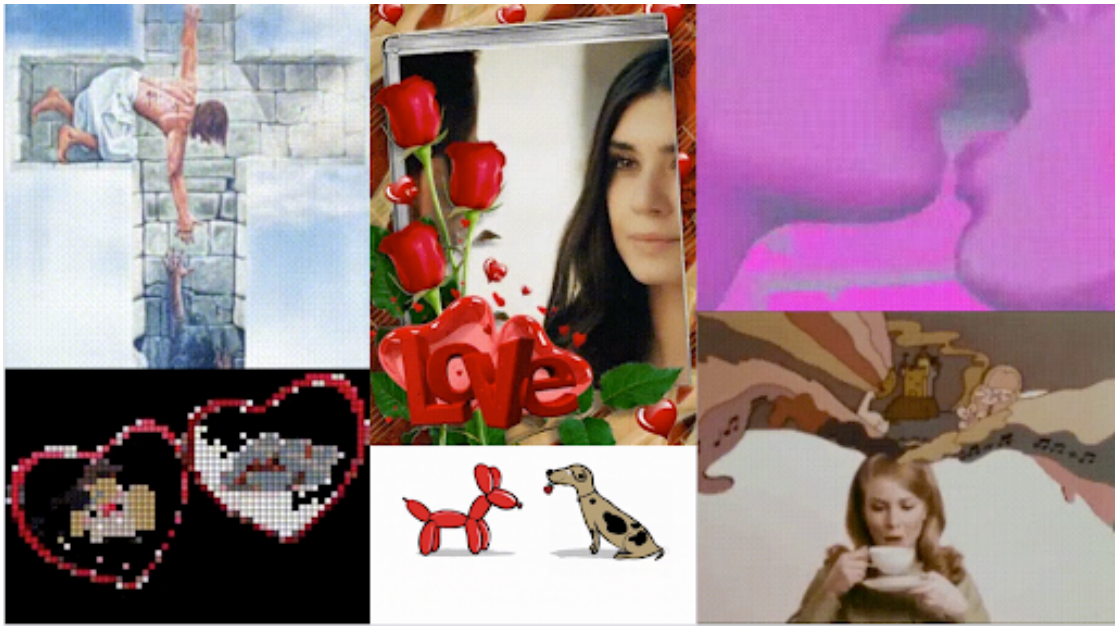
The remarkable success of the DL paradigm in CV over the past decade has given rise to more complex demands: methods to capture not only what an image "represents", but also what an image "evokes". This is a demand for tools that can accurately replicate the nuanced manner in which humans perceive the visual world, functioning at a "high semantic level" [174]. In essence, it represents an endeavor to move machine vision away from Funes' limitations, and towards genuine abstraction and reasoning capabilities. This emerging trend transcends the conventional focus of image classification on concrete classes, instead delving into the automation of intricate visual reasoning tasks deeply intertwined with subjective and cultural dimensions. These tasks encompass classifying images based on emotions [67, 257], discerning political affiliation [187], assessing beauty [149], and even inferring personality traits [321] solely from raw visual data. For humans, these tasks are heavily influenced by cultural contexts and biases. As such, the demand for the automation of these tasks has redefined the expectations placed upon CV models and the depth of knowledge they are tasked with acquiring.

I.1.1.3 ACs in the Era of Algorithmic Visual Curation

Abstract concepts (ACs) such as *comfort*, *freedom*, or *danger* are becoming important tools for advancing the next generation of automated visual indexing. These concepts, which underlie the human expression of emotions, opinions, and ideas through language [205], hold particular influence in categorizing and managing visual data. This is because visual forms such as photographs and paintings are thought to illustrate, and circulate, concepts both by providing links to depicted objects through raw features–such as lines, color, shape, and size–as well as through what Barthes called an image's *connotation*: a second layer of meaning made from culturally coded elements [33]. This indexing power of AC categories has been widely recognized in the domain of CH, where controlled thesauri and classification systems incorporate ACs to categorize visual materials [286]. Shared vocabularies and ontologies such as Iconclass,[1] Library of Congress,[2] and the Getty's Art and Architecture Thesaurus[3] offer pre-established ACs for association with visual content.

The effectiveness of ACs in the realm of visual indexing arises from their capability to bring together images that are visually diverse but semantically related (see Figure I.1.2). This is because, by definition, ACs lack distinct, concrete referents and are triggered by diverse scenarios [52], establishing a parasitic relationship between sensory-perceptual experiences and distributional linguistic data [101]. In the cognitive science field, the term "abstract concepts" [53, 160, 371, 392] is used for concepts that do not possess a single and perceptually bounded object as referent, and which have more variable content both within and across individuals than concrete concepts [29, 52]. The mechanisms that underlie the formation and use of ACs in human cognition are the object of study of the "Words As social Tools" (WAT) cognitive theory [52], which sees words as tools to perform action modifying the state of our social environment. In this sense, ACs are seen as tools that change the state of humans' inner processes, helping us formulate predictions and facilitating perception, categorization, and thought [52]. Advances in neuropsychology posit that, as opposed to concrete concepts, ACs rely on semantic rather than categorical similarity relations [92, 91], on associative relations [110], on the social, event, and introspective aspects of situations [31] and on evoked emotions [205, 370].

[1] https://iconclass.org. Access date: December 2023.
[2] https://www.loc.gov. Access date: December 2023.
[3] https://www.getty.edu/research/tools/vocabularies/aat. Access date: December 2023.



(a) Assortment of WhatsApp GIFs returned for the query *love*, illustrating the efficacy of ACs in visual indexing. ACs can unite visually diverse yet semantically related images, facilitating effective management of image collections.

(b) Assortment of Tate Gallery artworks returned for the query *consumerism*. Expert curators labeled each artwork with the *consumerism* tag. Identifying a unifying AC for these visually diverse artworks may challenge non-experts, highlighting the difficulty of establishing universally agreed-upon ACs in visual indexing.

Figure I.1.2: ACs bridge cultural connotations with visual forms.

I.1.2 AC Image Classification and Its Challenges

Given the pivotal role that ACs can play in the organization and categorization of visual data, and the current trend in automating image classification for increasingly abstract categories, the field of Abstract Concept-based image classification (hereafter referred to as *AC image classification*) is gaining substantial attention within computer vision, as will extensively be discussed in Chapter I.2. Within this dissertation, we formulate AC image classification as a multi-class classification problem. We opted for a single-label multi-class classification approach, rather than a multi-label multi-class approach, to focus on identifying the most prominent AC category for each image. This decision was made considering the challenge of reconciling disparities within visual representations of ACs. In many cases, images may not strictly belong to a single class, and ACs are not always mutually exclusive. By focusing on a single-label approach, we aim to capture the primary association between an image and its most salient AC. To clarify, the task involves working with a dataset comprising images $X = [I_1, I_2, \ldots, I_m]$, each paired with corresponding ground truth labels $Y = [y_1, y_2, \ldots, y_m]$. These labels are drawn from a set of K potential classes, where K encompasses a set of AC categories, such as [death, danger, comfort, ..., safety]. The primary objective is to ascertain the optimal image representation, I_i, and model parameters, θ, that enable us to predict the label $\hat{y_i}$ in such a way that it closely aligns with the true label y_i. This objective is succinctly expressed through the following equation:

$$\hat{y_i} = \arg\max(p(y_i|I_i, \theta)) \tag{I.1.1}$$

However, determining what the ideal image representation I_i should be, and emphasizing features that evoke specific ACs is complex–even for humans–due to the inherently elusive nature of these connections. In essence, identifying specific attributes that "objectively" classify a set of images as a certain AC is a challenge due to the diverse visual elements present (consider the examples in Figure I.1.2), even for human observers. Unsurprisingly, CV methods, while striving to replicate human perception, face similar challenges, as they rely on raw or concrete features. These challenges in AC image classification can be distilled into four critical factors.

Scarcity: Ground Truths, Baselines, Shared Methods

The CV community has only recently "made first attempts in tackling content which requires subjective judgment or abstract analysis" [174]. However, this progress faces significant hurdles, primarily stemming from the limited availability of AC ground truth labels in publicly accessible image datasets. This scarcity is compounded by class imbalances in widely used image datasets like ImageNet [105], which emphasize tangible object classes. The scarcity of well-defined abstract

ground truth annotations [1, 8] poses a formidable obstacle in effectively training data-hungry models. Furthermore, data scarcity not only restricts research and shared methodologies but also hinders the establishment of a well-defined task and baseline benchmarks for AC image classification.

Figure I.1.3: Two images sharing a significant amount of low-level features (colors, shapes), and depicted objects (policemen, gun, street, background trees), while conveying contrasting higher-level semantics. The image on the left suggests an AC like *friendship*, whereas the image on the right implies an AC such as *violence*.

Figure I.1.4: Two images sharing fewer low-level features, yet exhibiting similar higher-level semantics. Both scenes are likely to be associated with the AC *violence*.

(a) Two cultural images sharing lower-level features, but conveying contrasting high-level meanings. Left: Tsukioka Yoshitoshi's *Child Calms a Horse by Covering His Head with Her Jacket* (1875). Right: Yoshitoshi's *Tajima Seitaro Murders His Wife When She Refuses to Return to Him* (1875). Color woodblock prints, Herbert R. Cole Collection, Los Angeles County Museum of Art. Wikimedia Commons.

(b) Two cultural images sharing high-level semantics but less low-level features. Left: Artemisia Gentileschi's *Judith Beheading Holofernes* (1611-12), oil on canvas, 159 x 126 cm, Museo Nazionale di Capodimonte, Naples. Wikimedia Commons.

Figure I.1.5: Comparison of cultural images illustrating differences in feature sharing and high-level semantics.

Entropy: Going Beyond the Surface

While strides have been taken towards achieving comprehensive image comprehension in computer vision, with impressive performance in tasks such as object detection and image generation [25], this progression aligns with the deep learning paradigm's focus on low-level or concrete features, focused on tasks with perceptually bounded ground truths. Specifically, CV has had a strong focus on CNNs [217] and other DL methods that target images' low-level features, which have largely banked on the decontextualized nature of visual signals. However, the field's focus on CNNs and their reliance on decontextualized visual signals becomes problematic for high-level visual understanding tasks, because CNNs are tailored for high intra-class similarity situations [42, 328]. As such, CNNs' capabilities do not align well with the inherent diversity and nuanced cultural nature of ACs, which by definition lack distinctive bounded perceptual features (see Figures I.1.4 and I.1.3). This challenge applies not only to natural but also cultural images (see Figure I.1.5). Instead, abstract words are associated with higher dispersion ratings, resulting in an entropic, broader range of images returned from a query [200, 215]. This diversity in visual signals challenges the development of algorithms that can consistently recognize ACs from solely visual data. The wide variability in visual representation adds complexity to the training process and requires strategies to effectively identify the core features that characterize these concepts.

Subjectivity: Bridging Semantic and Cultural Gaps

The field of CV grapples with a fundamental challenge known as the *semantic gap*, defined by Smeulders as the "disparity between the information extracted from visual data and the interpretations it holds in specific contexts" [332, p. 1352]. In the realm of AC image classification, this semantic gap is further compounded by the diverse perspectives and cultural contexts of viewers. This diversity leads to ground truth data that is inherently unpredictable due to subjective and cultural biases: viewers' individual backgrounds and experiences give rise to distinct associations of ACs with images. For instance, a single image can elicit varying affective reactions based on cultural influences, personal attributes, and social circumstances [404]. In the context of Digital Humanities and computational visual studies, van Noord has introduced the concept of the *cultural gap* [267], which extends the semantic gap to encompass temporal and social dimensions. The cultural gap accentuates the dissonance between the visual data extracted by machine vision and the interpretations derived by various cultural groups, especially pronounced when dealing with polysemic images. The iconic 1989 *Tank Man* photograph (Figure I.1.6) provides a compelling example: it elicits distinct cultural interpretations, symbolizing *protest* in much of Western media, but military *re-*

straint according to Chinese officials, highlighting the diverse range of high-level meanings a single image can convey [267, p. 3].

Figure I.1.6: The iconic 1989 photograph portrays an unidentified civilian, nicknamed 'Tank Man,' standing in protest in Tiananmen Square, Beijing. This image exemplifies the cultural gap in visual interpretation, as it can be associated with different AC categories, such as *protest* and *restraint*, depending on the viewer's cultural perspective and biases. Source: The Associated Press, originally photographed by Jeff Widener.

Ethics: Stabilizing Unstable Meanings

The inherent cultural variability in how ACs are visually evoked not only introduces technical complexity but also raises ethical concerns about whose perspectives and values are being amplified through the automation of visual sensemaking. Recent ethical concerns surrounding cultural bias in Computer Vision (CV) pipelines have prompted the development of innovative frameworks aimed at transparently representing the knowledge assumptions within datasets and deployed AI systems, as discussed in works like [136, 252]. These concerns are particularly pronounced when addressing culturally biased labels applied to images featuring

human subjects, as demonstrated in studies such as [61, 60]. The complexity deepens when dealing with ACs as target classes, as these abstract and subjective categories are inherently shaped by social constructs and encompass a wide array of meanings influenced by cultural and historical contexts and power systems. For instance, labels such as *beauty*, *danger*, or *terrorism* applied to images of individuals are profoundly influenced by the context of annotation, reflecting the cultural perspectives of the annotators. The critical idea here is that AI systems are not neutral but instead power systems, which reflect the priorities, preferences, and prejudices of those who shape the models and the data they are trained on [256, 48, 19]. The impact of these labels goes beyond the technical realm and significantly affects individuals and communities [83]. When decontextualized labels are employed in training data, they can persist in the development of automated detection and classification systems. Utilizing AC image models without comprehending their decision-making processes can result in harmful outcomes, including the perpetuation of prejudice and racism.

I.1.3 Core Technical Concepts

Considering the challenges in high-level visual understanding tasks like AC image classification, we first center our research around three technical concepts:

I.1.3.1 Situated Grounding

The development of ethical AI necessitates transparency, accountability, and ethical considerations [87], especially for highly sensitive tasks like AC image classification. Throughout this work, I will define and develop the core concept of *situated ground truths*, aligned with Donna Haraway's exploration of *situated knowledges* [158]. It promotes ground truths rooted in complex, situated experiences rather than an assumed universal, objective standpoint. This effort is part of a greater trend in CV: given that ground truth datasets for CV research heavily reflect Western-centric perspectives, leading to inaccuracies when applied beyond this context, novel frameworks aim to explicitly represent underlying knowledge assumptions within datasets and deployed AI systems. Examples of such frameworks include data sheets [136], as well as model cards [252]. Notable scholarly works like [48, 256] adopt a decolonial perspective to critique AI, highlighting how these technologies unwittingly reinforce colonial power dynamics and values. Similarly, the authors of [19] challenge the traditional concept of a singular, definitive truth in human annotation, introducing the concept of *crowd truth* rooted in subjective human interpretation. Authors in [249] emphasize the importance of adopting a power-aware approach to data design and production for equitable outcomes.

This dissertation is guided by the idea of anchoring training data in a situated manner to enable AI to engage with information aligned with human perspectives, navigate complexities, and counteract biases stemming from assumed objectivity.

I.1.3.2 Performance

Performance is a key metric in AI and is used to evaluate new methods in CV, including aspects such as accuracy, efficiency, and robustness. As discussed above, for high-level visual understanding tasks, achieving elevated performance is a technical challenge in and of itself. Addressing this issue involves not only mitigating data scarcity but also acknowledging the influence of training dataset composition and format on model performance. Critically, the complexity inherent in achieving high performance in AC image classification paves the way for potential interdisciplinary collaborations. Fields such as cognitive science or linguistics can offer valuable insights into the processes that underlie human comprehension of visual data and ACs. In this context, the pursuit of enhanced technical performance in AC image detection may necessitate a greater emphasis on cognitive insights and neuro-symbolic techniques, as suggested by Hitzler [165], compared to approaches employed in more concrete visual tasks. Strategies such as informed learning with prior knowledge and reasoning on background knowledge for visual tasks, as proposed by Aditya et al. [4], show promise, leveraging the complementary strengths and weaknesses of statistical (data-driven) and symbolic (knowledge-driven) approaches to AI—a challenge and opportunity widely recognized in modern AI [38]. In this regard, incorporating insights from symbolic inference systems enriched with linguistic, sensory, perceptual, social, and cultural information can guide the learning process.

I.1.3.3 Interpretability

Interpretability is the third foundational concept of this work. The concept's emergence can be seen as a response to the perceived "explanatory deficit" within technical domains [44], driven by challenges posed by opaque deep models and the recognition of biases that can lead to inequities. The subfield of eXplainable AI (XAI) advocates for interpretability as a means to address these issues [254]. This becomes particularly crucial in the context of AC image classification, as it heavily relies on subjective, cultural, and interoceptive processes. The implications of reusing models without explainability are substantial, as it can inadvertently perpetuate harmful stereotypes or biased worldviews [272]. Consider, for example, a CV system that boasts high performance in classifying images with the AC *freedom*. While high-performance metrics like accuracy might seem impressive,

their value diminishes if there is no insight into how the system makes its predictions. The integration of interpretability into AI models not only reveals biases but also aids in their correction, ensuring ethical alignment and enhanced performance. Balancing model complexity with interpretability is a key challenge in explainability. Complex models often excel in performance but lack transparency. To address this, techniques like attention mechanisms, feature visualization, and rule-based explanations can offer insights into decision-making. Another promising approach, similar to performance enhancement, involves integrating hybrid symbolic-statistical systems. These systems combine semantic models with symbolic reasoning, providing valid justifications for input/output correlations, addressing the 'black-box' problem in modern machine learning systems [380].

I.1.4 Research Focus and Objectives

Guided by the AC image classification's challenges and the three core technical concepts, this dissertation is structured around three central research questions:

RQ 1: *To what extent can the end-to-end DL paradigm, connecting raw pixel values directly to unsituated AC labels, address the task of AC image classification in terms of both performance and interpretability?*

RQ 2: *Is it possible to automatically identify intermediary semantic features to bridge the semantic gap between raw pixels and ACs? How might the utilization of these features impact the performance and interpretability of AC image classification?*

RQ 3: *Is it possible to enhance the performance and interpretability of AC image classification by combining perceptual information with cultural and commonsense knowledge?*

These research questions form the basis of the dissertation's structure, which addresses specific challenges across three main areas:

- **Deep Learning for Abstract Concept Image Classification** (RQ 1): Investigating the effectiveness of state-of-the-art DL models in handling ACs in image data through an end-to-end approach.
- **Minding the Gap with Cognitive Intermediaries** (RQ 2): Exploring the potential of visual data descriptors to bridge the gap between raw pixels and ACs via perceptual semantics.
- **Reifying and Reasoning with Knowledge Graphs** (RQ 3): Analyzing the potential of semantic technologies in representing the commonsense and cultural dimensions to enhance AC image classification.

I.1.5 Structure of Contributions

This dissertation comprises five parts: Part I offers essential research background, while Parts II, III, and IV delve into each primary research area. Part V serves as a comprehensive conclusion.

Part I: Background

This section establishes the foundational background for the dissertation, focusing on high-level visual sensemaking and the role of Abstract Concepts (ACs), including their definitions and cognitive foundations. Chapter I.2 lays essential groundwork by surveying human high-level visual sensemaking in conjunction with the aspiration of computer vision (CV) to automate this process. It explores the natural semantic hierarchy of images, from pixels to objects and higher-level concepts, identifying ACs as semantic units at the apex of this 'semantic pyramid.' The chapter delves into the challenges of automatically recognizing ACs in visual data, systematically reviewing CV research dedicated to bridging this gap, and offering a comprehensive overview of the state of the art in automatic detection of ACs from images.[4] Chapter I.3 provides a concise foundation on the nature of ACs, highlighting their diverse and nuanced characteristics from a cognitive science perspective. It introduces the cognitive concept of "acquired embodiment," explaining how abstract words become linked to sensory-motor information through their associations with concrete words, thereby enhancing their representation. Additionally, the chapter explores the cognitive substrates of AC representation, laying the groundwork for understanding how ACs are represented in the human brain and how this knowledge can be applied to AI applications in subsequent parts of the dissertation.

Part II: Defining + Benchmarking AC Image Classification

This part addresses Research Question 1, exploring two sub-research questions:

> **RQ.1.1** *Is it feasible to construct an image dataset following the principles of the end-to-end DL paradigm, where images are associated with AC labels?*

Chapter II.1 introduces ARTstract, a novel image dataset comprising over 14,000 cultural images labeled with ACs. The chapter details the dataset's creation process, including integration, composition, and statistics, highlighting class imbal-

[4] Under review for publishing at ACM Computing Surveys, https://dl.acm.org/journal/csur

ances among AC clusters. It also addresses the challenges of defining and capturing AC ground truths, considering contextual variations. ARTstract serves as a valuable resource for studying the relationship between visual content and ACs in computational visual studies, digital humanities, art history, and cognitive science.

> **RQ.1.2** *To what extent can the end-to-end DL paradigm address the task of AC image classification in terms of representation, performance, and explainability?*

Chapter II.2 examines the effectiveness of end-to-end DL in AC image classification. It focuses on bridging the semantic gap by directly processing raw pixel data and mapping it to AC target classes. This paradigm utilizes pre-trained DL models to transform raw images (I_{RAW}) into deep feature vectors (I_{DL}) used for classification, aiming to maximize class probabilities based on the transformed image representation and model parameters. The chapter is organized into three sections, each addressing specific research questions: representation, performance, and explainability. Representation explores intraclass similarity within deep representations, performance evaluates state-of-the-art deep models' effectiveness, and explainability delves into the interpretability of the deep models. This last section involves classification analysis, saliency maps, perceptual topology exploration, and introduces a novel explainability technique called SD-AM, offering insights into model decision-making and human-readable feature visualizations.[5]

Part III: Minding the Gap with Cognitive Intermediaries

This part addresses the second research question, through two sub-research questions:

> **RQ.2.1** *Is it possible to translate the cognitive science idea of acquired embodiment into a computational method for detecting perceptual semantics meaningful to AC visual evocation?*

Chapter III.1 investigates the translation of the cognitive concept of *acquired embodiment* into a computational method. This method aims to identify *perceptual semantics* (PS), concrete sensory features that can visually ground ACs within images in a data-driven manner. The chapter presents a proof of concept by experimenting with Tate Gallery artworks.[6]

[5] This work has been published as D.S. Martinez Pandiani et al. "Hypericons for interpretability: decoding abstract concepts in visual data." In: *International Journal of Digital Humanities* (2023), pp. 1–40.

[6] This work has been published as D. S. Martinez Pandiani and V. Presutti. "Automatic Modeling of Social Concepts Evoked by Art Images as Multimodal Frames." In: *Proceedings of the Workshops and Tutorials held at LDK 2021 co-located with the 3rd Language, Data and Knowledge Conference (LDK 2021)* (2021)

> **RQ.2.2** *Can cognitive-inspired perceptual features be effectively leveraged and employed to enhance image representations for improved performance and explainability in the context of AC image classification?*

Chapter III.2 explores the use of PS to bridge the semantic gap by automatically extracting and exploiting concrete semantic labels from images. This paradigm adopts a feature engineering approach that converts raw images (I_{RAW}) into perceptual semantic representations (I_{PS}) that explicitly correspond to the presence of more concrete PS. As a result, the I_{PS} representation is characterized by a more interpretable, symbolic foundation. We harness these image representations to train interpretable classical Machine Learning (ML) techniques, such as Naive Bayes, for AC image classification, maintaining performance comparable to CNNs while enhancing interpretability.

Part IV: Reifying and Reasoning with Knowledge Graphs

This part addresses the third research question, through three sub-research questions:

> **RQ.3.1** *Can the incorporation of ontology-based frameworks facilitate automatic reasoning over PS to establish high-level concepts from commonsense linguistic knowledge?*

Chapter IV.1 explores the possibility of automatically reasoning over the concrete semantics of visual data to establish connections with high-level concepts, emphasizing interpretability. A method is developed to reason over visual descriptors' textual labels and establish connections with linguistic frames using commonsense knowledge. This work enhances multimodal sensemaking by establishing interpretable connections between images and high-level conceptual frames via ontology-based reasoning.[7]

> **RQ.3.2** *Do PS serve as entry points for subjective and cultural biases in their assignment to visual data, and if so, what effective methods can be employed for representing and addressing these biases?*

Chapter IV.2 focuses on capturing subjectivity and cultural bias in data labeling in a machine-readable way for training AI systems. A novel ontology-based method, SituAnnotate, is introduced for situating ground truths in the context

[7] This work has been published as Fiorela Ciroku et al. "Automated multimodal sensemaking: Ontology-based integration of linguistic frames and visual data." In: *Computers in Human Behavior* 150 (2024), p. 107997. Author order followed alphabetical arrangement.

of their annotation. The chapter discusses challenges, approaches, provides selection guidelines, and presents real-world use cases with the case study of image annotations.[8]

> **RQ.3.3** *To what extent can the grounding of perceptual features within their subjective, cultural, and commonsense contexts be leveraged to enhance AC image classification in performance and interpretability?*

Chapter IV.3 explores the use of Knowledge Graphs (KGs) to enhance the task of AC image classification via the integration of perceptual semantics with cultural and commonsense symbolic knowledge. The chapter introduces the ARTstract-KG, which expands traditional image datasets by providing PS labels, contextual annotations for all images, and pre-reasoned commonsense knowledge. The chapter transforms PS representations of images (I_{PS}) into structured KG format (I_{KG}), linking images to perceptual, situational, and commonsense semantics. Experiments involve embedding of the KG into vector space, resulting in vector representations of images ($I_{\text{PS-KGE}}$) based on situated KG information. These representations are utilized in the AC image classification task. Additionally, hybrid approaches are explored to bridge the gap between the end-to-end deep vision paradigm and the situated perceptual knowledge paradigm. This approach ensures compatibility between representations, combining deep embodied features (I_{DL}) with embeddings from situated KG information (I_{KGE}) to create an integrated representation (I_H). The chapter explores both absolute and relative representations [261] of these embeddings for use in the AC image classification task. Our results outperform the state of the art and offer enhanced interpretability.

Part V: Towards Hybrid Cognitive-Inspired AI

Chapter V.1 synthesizes the findings of the research, addresses the research questions, highlights significant contributions, identifies remaining inquiries and future research directions, and provides reflective insights to guide future work.

[8] Parts of the chapter have been published as D.S. Martinez Pandiani and V. Presutti. "Coded Visions: Addressing Cultural Bias in Image Annotation Systems with the Descriptions and Situations Ontology Design Pattern." In: *6th International Conference of Graphs and Networks in the Humanities 2022: Technologies, Models, Analyses, and Visualizations* (2022), while other parts are in revision for publishing as D.S. Martinez Pandiani and V. Presutti. "Situated Ground Truths: Enhancing Bias-Aware AI by Situating Data Labels with SituAnnotate." In: *Special Issue on Trustworthy Artificial Intelligence of ACM Transactions on Knowledge Discovery from Data (TKDD)* (2024).

Chapter I.2

Seeing the Intangible: A Survey

Summary The field of Computer Vision (CV) was established with the ambitious goal of achieving *complete image understanding*, entailing a complete semantic interpretation of input images. However, the exact nature of this goal remains elusive. In human visual understanding hierarchies, there exists a top-level category referred to as "high-level" semantics, housing the most intricate and subjective information discernible from visual data. Within the CV field, there's a growing emphasis on classification and detection tasks focused on these "high-level semantics," albeit lacking a precise definition. Consequently, recent CV research has introduced varied terminologies describing similar non-concrete semantic elements, many of which represent Abstract Concepts (ACs). These ACs play a pivotal role in high-level visual sensemaking and thus their significance extends to advanced image management and retrieval processes. This survey systematically reviews CV research associating still images with high-level semantic units, particularly focusing on tasks dealing with ACs. We first establish a clear characterization of "high-level" semantics in human image understanding. Subsequently, we identify tasks, datasets, and approaches associated with these high-level semantics. We categorize the various CV tasks operating at this level and provide a roadmap for future research to define and refine their utilization of the term "high-level visual semantics." We also conduct a systematic, in-depth review of research addressing tasks most closely resembling AC-based image classification. This work highlights a growing focus on culturally-dependent labels in CV, emphasizing the importance of task- and context-specific datasets and the field's evolving capability to handle abstract semantics in visual data. Notably, it suggests that accumulating extensive datasets does not necessarily guarantee high F1 scores, underscoring the need for more sophisticated approaches.

I.2.1 Introduction

Visual imagery has historically been a potent medium for conveying both abstract and concrete ideas, a significance evident in the vast amount of images shared daily on social media [113]. This surge in visual content has fueled extensive research in CV, primarily aimed at automating the indexing, retrieval, and management of visual data, with applications spanning disciplines like sociology, media studies, and psychology [187, 17]. CV's data-driven approach, treating images as data, has been pivotal, facilitated further by the recent deep learning (DL) paradigm shift, leading to significant achievements in tasks such as image classification, object detection, and image generation [25]. The remarkable success of the DL paradigm in CV has led to more intricate demands, including the need for tools capable of replicating human-like perception at a "high semantic level" [174]. This includes using CV to classify images based on high-level notions, known as Abstract Concepts (ACs), which have proven instrumental in various tasks such as emotion classification [67, 257], political affiliation detection [187], beauty assessment [149], and personality trait inference [321], all accomplished through raw visual data.

However, explicit definitions of high-level visual semantics, particularly ACs, in machine vision are sparse. This lack of clarity, combined with the historical emphasis on physical object detection grounded in low-level feature analysis, often results in less impressive results in high-level semantic tasks compared to concrete object classes [52]. Additionally, these tasks are influenced by cultural contexts and human biases in perception, which redefine the depth of knowledge and understanding expected from CV models. Our survey systematically reviews CV studies addressing the challenge of automatically classifying visual data based on high-level semantic units. We clarify what constitutes "abstract" or "high-level" semantics in the context of an image and identify CV tasks and automatic detection approaches related to these semantics. Focusing on abstract concept-based image classification (AC image classification), particularly in still images, we conduct a comprehensive overview of the state of the art. This includes:

1. **High-Level Semantic Units**: Identification and clustering of high-level semantic units, integrating insights from cognitive science, visual studies, art history, and computer science.

2. **High-Level CV Tasks**: Surveying of the CV landscape to identify and cluster tasks associated with high-level visual sensemaking, while examining common methodologies and datasets.

3. **AC Image Classification**: We conduct a detailed review of works dealing explicitly with AC image classification in still images.

This chapter is structured as follows. Section I.2.2 provides an interdisciplinary examination and characterization of what constitutes "full" or "high-level" semantics in human visual understanding. In Section I.2.3, the methodology employed to identify works related to the high-level semantics in the CV field is described. Section I.2.4 surveys and categorizes CV tasks and works associated with high-level visual understanding, facilitating the discovery of implicit CV research addressing ACs. In Section I.2.5 we perform a thorough survey of CV-based works that research tasks analogous to AC image classification. Section I.2.6 presents datasets potentially relevant to the AC image classification task. The implications and contributions of the survey are discussed in Section I.2.7. Ultimately, Section I.2.8 provides concluding remarks. More details are available and documented in a specialized GitHub repository.[1]

I.2.2 Defining High-Level Visual Semantics

I.2.2.1 Three-Tiered Semantics

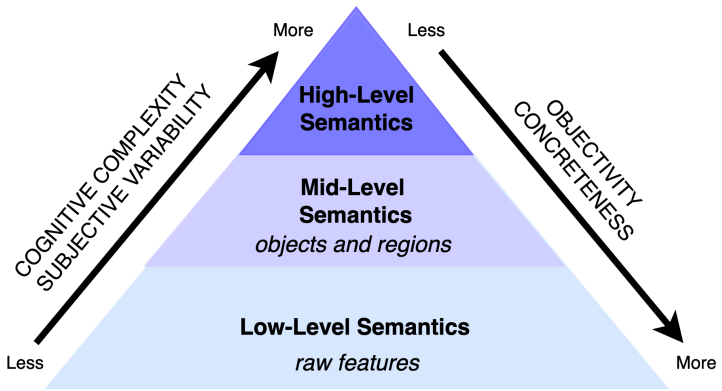


Figure I.2.1: The three tiers of the visual semantics hierarchy. Visual understanding is often depicted as a multi-layered process, revealing three distinct levels of semantics. The low level involves raw or elemental features, while the mid-level encompasses individual objects, persons, and regions. In contrast, the high level remains less defined and explored.

The concept that the perception and interpretation of visual meaning involve a

[1] https://github.com/delfimpandiani/seeing_the_intangible. Access date: February 2024.

multi-layered process is a shared perspective across various domains and applications, including cognitive science, CV, content-based image retrieval (CBIR), and visual studies. This multi-layered nature was emphasized in the seminal paper by Hare et al. (2006) [159], which discussed Smeulders' idea of the "semantic gap" in CV [332]. This paper also highlighted the common practice of referring to different strata of meaning within images, a concept that has been pivotal in CBIR. We delved deeper into several of these multi-layered approaches, drawing insights from works by Panofsky (1955) [278], Shatford (1986) [327], Greisdorf and O'Connor (2002) [150], Eakins (2000) [112], Jorgensen (2003) [188], Hare et al. (2006) [159], and Aditya et al. (2019) [4]. This exploration revealed a general analogy wherein three semantic tiers are used to delineate the human visual understanding process: a "low-level," a "mid-level," and an "upper-" or "high-level" tier, corresponding to increasing complexity, variability, and subjectivity (see Figure I.2.1). Most of these approaches represented these layers using a pyramid analogy to illustrate a hierarchical structure. Via a thorough analysis of the semantic elements assigned to each of the layers by each of the foundational works, we noted that there was a consensus in identifying and agreeing upon semantic units within both the low- and mid-level layers. However, this consensus did not extend to the topmost layer.

At the base, the "low-level" layer (depicted in light blue in Figure I.2.1) encompasses raw or primitive features such as regions, edges, textures, colors, shapes, and textures. Moving up to the "mid-level" layer (depicted in light purple in Figure I.2.1), this tier accommodates semantic entities like objects, persons, regions, and places. Much of CV research has centered on this layer, emphasizing object recognition and image segmentation. In contrast, the "high-level" layer of semantics (depicted in dark purple in Figure I.2.1) remains less detailed and subject to less consensus. This topmost tier, often associated with the concept of "full semantics," lacks an explicit and consistent definition, and characterization of what types of semantic units belong in it. Instead, there appears to be a *tacit* shared understanding of the kinds of content that may reside or be conceived within this layer. In our analysis, this layer emerged as both elusive and significant, akin to the "tip of an iceberg" regarding visual semantics, motivating our efforts to define it more precisely.

I.2.2.2 Tip of the Iceberg: Upper Visual Semantics

Images may be sought "on the basis of their holistic content or message, as opposed to the information embedded within them by dint of their depiction of certain features" [115, p. 39]. Most work that attempts to name and characterize where and how such holistic content arises thus moves in a layered way further away from raw or primitive features, to arrive at the "highest" tier of the semantic pyramid, referred to with different names: iconological layer [278], higher level of

understanding [188], abstract content [327], abstract attributes [112], subjective beliefs [150], higher level semantics [4], or full semantics [159].

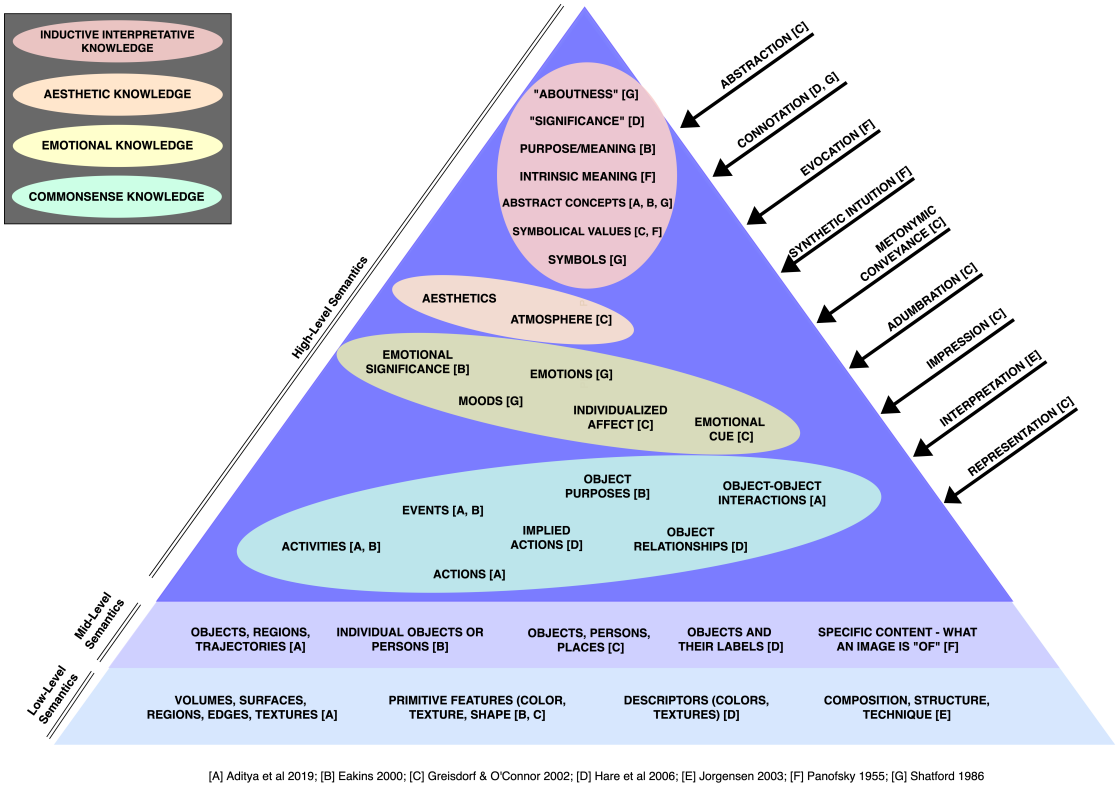


Figure I.2.2: Tip of the iceberg: a deeper characterization of the top level of the visual semantic pyramid. Drawing from a multidisciplinary exploration of semantic entities associated with this upper semantic layer, we have identified four distinct clusters of knowledge.

Part of the difficulty of solidifying a cross-disciplinary shared understanding of high-level semantics is that, in comparison to the other levels, high-level understanding by humans is increasingly cognitively complex. Complex cognitive processes, including abstraction, metonymic conveyance, adumbration, impression, prototypical displacement [150], connotation [159, 327], evocation, and synthetic intuition [278] are considered crucial tools for understanding visual semantics at this "high level" of abstraction. However, it is generally thought that it is practically hard to grasp them using typical automatic image understanding and indexing methods. As such, this highest level of abstraction in the interpretation of image meaning or content is seen as a "seemingly insurmountable obstacle" to the application of content-based image retrieval techniques [115].

In addition to cognitive complexity, subjectivity represents another challenging

aspect when it comes to characterizing and automatically recognizing semantic units within this level. Shatford's widely cited insight encapsulates this notion succinctly: *"...the delight and frustration of pictorial resources is that a picture can mean different things to different people"* [327, p. 42]. Furthermore, a single picture can convey diverse meanings not only to various people but also to the same individual in different contexts or at different times of necessity. In line with this perspective, Greisdorf [150] underscores the importance of interdisciplinary perspectives as a foundational approach for modeling the attributes of the human image cataloging process, because:

> Those attributes tend to elude the indexing/cataloging process by exceeding the image indexing threshold due to individual viewer cognitive displacement of objects and object features that give rise to disjunctive prototypes that viewers may associate with the objects included as part of the image composition. These adumbrative, impressionistic and abstractionist concepts that relate viewer to image need to be captured with some type of retrieval mechanism in order to enhance retrieval effectiveness for system users. [150, p. 11]

To better comprehend and communicate about these abstract semantics, there is a need to precisely identify the semantic units that may belong to this layer and potentially characterize their interrelationships. Thus, we systematically reviewed the cited literature to provide a more detailed characterization of this apex of visual semantics (see Figure I.2.2). We categorized the types of elements mentioned as belonging to high-level visual semantics into four general groups:

- **Commonsense semantics.** This cluster is closely aligned with mid-level semantics and is among the least subjective of the groups. It encompasses semantic elements such as explicit or implied actions ("running") [4, 159], activities ("dance") [4, 112], events ("parade") [4, 112], relationships between objects or object-object interactions ("man holding cup") [4, 159], and object purposes [112]. These elements fall under the category of "commonsense" knowledge because they often exhibit a high level of consensus among viewers and can be described within a logical framework.

- **Emotional semantics.** This cluster encompasses semantic information associated with emotions, encompassing moods, emotions [327], emotional cues [150], emotional significance [112], and individualized affects [150].

- **Aesthetic semantics.** This smaller cluster focuses on global aesthetic attributes that pertain to the overall judgment of an image as a unified entity. Semantic units within this cluster include atmosphere [150].

- **Inductive interpretative semantics.** Positioned as the uppermost cluster, this group contains some of the most complex, subjective, and culturally encoded semantic elements within "high-level semantics." It encompasses semantic units like an image's "aboutness" [327], significance [159], purpose, and meaning [112], including intrinsic meaning [278]. Crucially, this cluster includes symbols [327], symbolical values [150, 278], and abstract concepts [4, 112, 327] as part of the top tier of visual semantics.

These clusters represent our initial effort to provide a preliminary characterization of the high-level layer of visual semantics. Although they are presented as distinct categories, there may be instances where semantic elements overlap between clusters, such as the intersection of mood and aesthetics, atmosphere and ACs, or emotion and ACs. Ongoing work may lead to further refinements and revisions of this diagram.

I.2.2.3 Abstract Concepts and Visual Data

In the cognitive science field, the term abstract concepts (ACs) [53, 160, 371, 392], e.g. *violence, freedom* or, *danger*, refer to complex situations which do not possess a single and perceptually bounded object as referent, and which have more variable content both within and across individuals than concrete concepts [29, 52]. The mechanisms that underlie the formation and use of ACs are the object of study of the "Words As social Tools" (WAT) cognitive theory [52, 54], which sees words as tools to perform actions modifying the state of our social environment. In this sense, ACs are seen as tools that change the state of humans' inner processes, helping us formulate predictions and facilitating perception, categorization, and thought [52].

As such, ACs are valuable for tasks like automatic indexing, retrieval, and managing visual data. Enser recognized over 20 years ago that concept-based image retrieval methods would continue to be vital for archival image collections [115, p. 206]:

> At the highest level of abstraction in the interpretation of image meaning or content, i.e. that which corresponds with Panofsky's iconological level, the human reasoning based on tacit or world knowledge which underpins image indexing and retrieval operations poses a seemingly insurmountable obstacle to the application of CBIR techniques. At this level, we humans are able to 'see' within the primitive attributes of two dimensional imagery the portrayal of *love, power, benevolence, hardship, discrimination, triumph, persecution and a host of other aspects of the human condition.* We are enabled, through the visual medium,

> to exercise skills in semiological analysis – the shared connotation of the icon, metonym and metaphor, the understanding and appreciation of two conceptually related but antithetical images

Importantly, the automatic association of ACs to images could lead to breakthroughs in a wide range of applications, including Web image search, online picture-sharing communities, and scientific and academic endeavors. It would benefit multiple applications by improving multimedia querying of digital libraries, as automatically generated ACs could be used for Boolean search, as traditional metadata are. Furthermore, enabling machines to recognize the potential of images to communicate ACs could be useful to CH institutions, to enhance their visual collections' documentation, descriptions, and mediation–for example, using detected ACs to inform the design of multimodal interactive environments. Additionally, it could benefit public institutions building narratives about their visual objects to engage people from different backgrounds or with different abilities, companies in the creative sector exploiting their already existing catalogues, and companies building products or services related to specific abstract concepts.

Despite the considerable potential of ACs as descriptors of 'aspects of the human condition' for visual indexing, the CV community has only recently started to tackle subjective and abstract content analysis [174]. At first glance, challenges associated with high-level visual semantics become evident, including subjective perception influenced by personal and situational factors [404], a lack of shared methods and communication among researchers, class imbalances in popular datasets, and increased variability in query results involving abstract words [1, 200, 215]. Furthermore, there is no explicit task definition or specific purpose datasets for AC image classification or AC detection from images. This lack of clarity necessitates an additional aspect of this survey: to explore past CV research that may have addressed the task implicitly, using different terminology.

I.2.3 Survey Methodology

The primary objective of this survey is to identify CV methodologies, tools, and architectures designed for associating high-level semantic units, including ACs, with still images. In this section, we will outline our selection criteria for choosing papers within the CV field that address high-level abstract visual understanding tasks. This includes the criteria applied to select relevant publication venues and the keywords used for mining these repositories and venues. We will also elucidate the process for sub-selecting works that underwent additional detailed analysis for insights into AC image classification, and the criteria employed in the categorization of such works that implicitly or explicitly relate to the task.

Venue Selection Criteria Our venue selection was guided by two key factors: relevance and impact. We initially narrowed our search to venues focused on CV, aligning with our research area's primary focus. Additionally, we assessed venue impact using established criteria to ensure high-quality and significant selections, thus confirming their suitability for our research. For journals, we initially exclusively considered those with a Q1 rating in the field of Computer Vision and Pattern Recognition, as determined by Scimago[2]. This criterion resulted in the identification of 19 journals and one book series. Notably, we excluded journals related to the medical field,[3] leaving us with 15 selected journal venues. For conferences, we exclusively considered those with an A* or A rating according to CORE[4]. This selection process yielded four conferences that met our impact criteria. The complete list of the 19 selected venues can be found in the Github repository.

Category	Keywords
"Abstract" Adjectives (synonyms, hyponyms of, or similar to "abstract")	*abstract, intangible, non-concrete, non-physical, symbolic, latent, evoked, implied, subjective, social, cultural, moral, political, economic, affective*
"Concept" Nouns (synonyms, hyponyms of, or similar to "concept")	*concept, object, idea, class, value, ideology, emotion, affect, sentiment, signal, attribute*
CV Task Nouns (synonyms, hyponyms of, or similar to "detection")	*detection, classification, recognition, identification*

Table I.2.1: Keywords for query construction, categorized into adjectives similar or synonymous with 'abstract' and 'high-level', nouns denoting similar ontological objects to 'concepts', and task nouns. This categorization enables the creation of queries to identify works related to high-level concept detection or similar tasks.

Keywords and Query Building. High-level visual sensemaking within CV lacks standardized nomenclature. The preliminary investigation into high-level semantic units in Section I.2.2 revealed the interchangeable use of various terms to describe abstract notions across multiple disciplines. Subsequently, we compiled a list of keywords, considering the diverse terminology used to denote these concepts and

[2] https://www.scimagojr.com/. Access date: January 2021.

[3] Journal of the Optical Society of America A: Optics and Image Science and Vision, International Journal of Computer Assisted Radiology and Surgery, Medical Image Analysis, and Computerized Medical Imaging and Graphics

[4] http://portal.core.edu.au/conf-ranks/. Access date: January 2021.

high-level abstract understanding across different fields. Our search on major academic databases for publications related to AC image classification from a CV perspective informed the final list of keywords. These keywords are categorized into adjectives, nouns, and nouns describing related tasks like detection or identification. By combining terms from these categories, we aimed to construct queries capable of identifying works related to AC image classification or similar tasks (see Table I.2.1).

High Level Visual Understanding Search. The list of keywords was employed to construct search queries for the selected venues, with the inclusion of the term "image." All searches across the chosen venues were conducted using the SCOPUS[5] citation database. An exception was made for the ECCV Conference and Workshop, as they were not accessible through this database. The results were manually reviewed to collect papers specifically addressing high-level visual understanding of still images, i.e., those dealing with some of the semantic elements identified in Section I.2.2. Through this process, 52 papers were gathered, comprising 6 journal papers (2017–2021) and 44 conference papers (2008–2021). Subsequently, a pruning phase was conducted to exclude articles meeting any of the following criteria:

- Articles not written in English; only articles in the English language were considered for this survey.
- Articles describing high-level visual semantics tasks applied exclusively to moving images (e.g., videos), such as [108] and [93]. However, articles that applied methods to both video and still images were considered.
- Articles addressing image captioning in a general context, without a focus on high-level reasoning tasks (e.g., [219]).
- Articles describing methods that relied on additional data sources beyond raw pixel data, particularly those dealing with multimodal or cross-modal data (e.g., text and images from news articles together, etc.), as in [86], [356], and [362].
- Articles centered on image generation (e.g., [288]).
- Articles dealing with high-level visual semantics tasks, such as cultural event recognition, but produced within the context of the Chalearn LAP challenge (e.g., [229], [302], [376], [379]), as these have been previously surveyed [118].

[5] https://www.scopus.com. Access date: May 2022.

General Survey Criteria: This resulted in the selection of a total of 38 publications, comprising 4 journal papers and 34 conference papers. In our survey of these 38 publications, we aim to identify several key criteria, including:

- **Image Type:** We examine whether the publications focus on natural photographs, art images, or other image types.
- **Related Semantic Elements:** We identify the high-level semantic units of interest, including examples, in each of the publications.
- **Datasets:** We analyze the datasets used and whether the authors created them for their research.
- **Computer Vision Tasks:** We categorize the explicit CV-related tasks addressed in each publication.
- **CV Task Clusters:** We classify each publication into relevant clusters of CV tasks.

Subselection for In-Depth Surveying: Upon reviewing the initial set of 38 publications, we observed a wide spectrum of high-level CV tasks with broad coverage of semantic units. To align with our core objective of studying ACs more closely aligned with values and ideologies, we then did an additional round of surveying for works addressing semantic units referred to as "symbols" [174, 390, 391, 191], "intent" [183], and "abstract topics" [354, 355]. We did not consider works dealing with emotions, given the existence of a substantial number of surveys focused on them (e.g., [276]). By narrowing our selection, we ensured methodological alignment with our research objectives. With a relatively limited number of qualifying publications, we extended our exploration using a bottom-up approach. By meticulously reviewing the bibliographic references of previously surveyed works, we identified additional publications, including interconnected or extended works. Notably, pairs like [391] extending [390] and [355] extending [354] were considered together in our in-depth survey.

Specific In-Depth Criteria. For the 8 selected works, we conducted a more detailed analysis, classifying them based on various dimensions. In addition to the general dimensions, we introduced other criteria for our comprehensive survey:

- **Explicit vs. Analogous/Related Task**: Explicit task addressed versus analogous AC image classification task.
- **Model Architecture**: Backbone architecture for the analogous task (e.g., CNN with perturbation, transformer with KG and GCN).

- **Reproducibility**: Availability of software and data.
- **Performance Metric**: Reported macro F1 score for the analogous task, if available.
- **System Hybridity**: Classification of the system as statistical, symbolic, or hybrid based on [38]'s taxonomy.

I.2.4 Automatic High-Level Visual CV Tasks

Our survey aimed to evaluate the contemporary advancements in CV for automatically detecting high-level semantic units in static images. Following the outlined methodology in Section I.2.3, we extensively searched top-rated publications in Computer Vision and Pattern Recognition. This endeavor yielded 38 notable publications, comprising 4 journal papers (2017-2021) and 34 conference articles (2008-2021). Table I.2.4 presents an overview of our examination of selected works, aiming to uncover significant trends in the application of CV technologies for high-level visual semantic understanding. We explored various facets of these works, including their targeted image domains (e.g., natural photographs, art images), dataset utilization (and curation), and specific CV tasks addressed. Our primary focus was to identify the high-level semantic units addressed by each paper concerning static images, with illustrative examples provided where applicable. This thorough analysis provided valuable insights into the exploration of Abstract Concepts (ACs) within the existing literature.

Based on the table, distinct trends emerge, with a significant portion of this research gravitating toward specific CV tasks. These tasks include situation recognition [389, 344, 280, 222], social relation recognition [401, 377, 346, 223, 220, 147], event recognition [388, 395, 56], visual persuasion and intent analysis [187, 183, 172, 154], automatic advertisement understanding [391, 390, 174], visual sentiment analysis [388, 358, 2], aesthetic evaluations [385, 149, 98], and analysis of social and personality traits [321, 244, 186], including occupation [324], fashion [199, 171], and group-level analyses [143, 127]. Other CV tasks include abstract reasoning [337], affordance reasoning [81], political bias detection [355], visual humor detection [72], and environmental variables prediction [198]. These task trends seem to align with the identified semantic units in Section I.2.2, encompassing emotions, events, aesthetics, atmosphere, object purposes, object-object interactions, symbols, and more, reflecting the diverse landscape of high-level visual semantic exploration.

Year	Work	Image Type	Dataset	Own Data	Task(s)	Related Semantic Unit	Semantic Unit Examples	CV Cluster(s)
2006	[98]	natural photograph	Untitled	Y	aesthetic quality inference	aesthetics	*aesthetic value*	aesthetic analysis
2009	[127]	natural photographs of groups of people	The Images of Groups Dataset	Y	demographic information detection; event recognition	events	*dining, age, gender*	social signal processing; situational analysis
2010	[374]	natural photographs of groups of people	The Images of Groups Dataset	N	social relationship recognition	object-object interaction	*siblings, mother-child, husband-wife*	social signal processing; situational analysis
2010	[149]	natural photographs of frontal female faces	Untitled	Y	female facial beauty prediction	aesthetics; abstract concept	*female beauty*	aesthetic analysis; social signal processing
2012	[244]	natural photographs of faces	Untitled	Y	social dimension trait detection	individualized affect; object purposes	*dominance, aggressiveness, threatening*	social signal processing
2013	[324]	natural photograph	Occupation Database	Y	occupation prediction	object purposes	*waiter, customer, clergy*	social signal processing
2013	[56]	natural photograph	PEC Data Set	Y	event recognition	social events	*roadtrip, halloween, christmas*	situational analysis
2014	[187]	natural photographs of politicians	Visual Persuasion Dataset	Y	communicative intent recognition; visual persuasion understanding	object purposes; purpose/meaning	*trustworthy, powerful, energetic*	visual rhetorical analysis
2014	[198]	natural street view images	Untitled	Y	environment navigation	situational analysis; atmosphere	*proximity to a McDonald's, crime rate*	situational analysis
2014	[199]	natural photographs of full body outfits	Untitled	Y	fashion style classification	aesthetics	*hipster, bohemian, preppy*	social signal processing
2015	[186]	natural photographs of politicians	Visual Persuasion Dataset	Y	political affiliation detection; personality trait detection	individualized affect; object purposes	*intelligence, competence, democrat affiliation*	social signal processing; visual rhetorical analysis
2015	[395]	natural photographs of events	WIDER	Y	event recognition	event	*parade, press conference, meeting*	event recognition
2015	[403]	natural photographs with visible faces	Social Relation Dataset	Y	social relationship recognition	object-object interaction	*friendliness, warmth, attachment*	social signal processing; situational analysis
2016	[389]	natural photograph	Situnet	Y	situation recognition	actions; implied actions; object purposes; object relationships	*obstacle, audience, source*	situation recognition
2016	[72]	clipart images	AVH (Abstract Visual Humor) Dataset	Y	visual humor prediction	atmosphere; meaning	*humor/funninness*	visual rhetorical analysis
2016	[172]	natural photographs of politicians	Visual Persuasion Dataset	N	communicative intent recognition; visual persuasion understanding	object purposes; purpose/meaning	*trustworthy, powerful, energetic*	visual rhetorical analysis

Year	Work	Image Type	Dataset	Own Data	Task(s)	Related Semantic Unit	Semantic Unit Examples	CV Cluster(s)
2017	321	natural photograph	PsychoFlickr corpus	N	personality trait-based image classification	individualized affect; significance	*neuroticism, extraversion, openness*	social signal processing
2017	337	artificial chessboard images	Untitled	Y	symmetry task, identity task	object-object interaction	*identity, symmetry*	situational analysis
2017	174	image advertisements	Ads Dataset	Y	automatic advertisement understanding; symbolism prediction	symbols; symbolical values	*danger, beauty, death*	visual rhetorical analysis
2017	171	natural photographs (fashion images)	Style Embedding Dataset	Y	style embedding learning	aesthetics	*hipster, bohemian, preppy*	social signal processing; aesthetic analysis
2017	220	natural photograph	PISC	Y	social relationship recognition	object-object interaction; atmosphere	*friends, couple, professional relationship*	social signal processing
2017	346	natural photograph	PIPA-relation	Y	social relationship recognition	object-object interaction; atmosphere	*friends, colleagues, lovers*	social signal processing
2017	385	ground-level and overhead natural photographs	ScenicOrNot (SoN) dataset	Y	scenicness prediction	atmosphere; aesthetics	*scenicness of natural landscapes*	aesthetic analysis
2017	222	natural photograph	imSitu	N	situation recognition	actions; implied actions; object purposes; object relationships	*obstacle, victim, destination*	situation recognition
2018	81	natural photographs of scenes	ADE-Affordance	Y	affordance reasoning in scenes	object purposes	*socially awkward affordance, socially forbidden affordance*	situation recognition
2018	390	image advertisements	Ads Dataset	Y	automatic advertisement understanding	symbols; symbolical values	*freedom, happiness, adventure*	visual rhetorical analysis; situational analysis
2019	388	natural photographs, art images and art paintings	Flickr and Instagram (FI); IAPSa; ArtPhoto; Abstract	N	affective image retrieval	emotion	*contentment, awe, excitement, disgust*	visual sentiment analysis
2019	344	natural photograph	imSitu	N	situation recognition, role assignment	actions; implied actions; object purposes; object relationships	*audience, obstacle, source*	situation recognition
2019	143	natural photographs of groups of people	GAF-personage database	Y	most influential person identification	affect; object-object interactions	*leader, influential person*	social signal processing
2019	391	image advertisements	Ads Dataset	Y	automatic advertisement understanding	symbols; emotions; abstract concepts	*danger, adventure, death*	visual rhetorical analysis
2019	147	natural photograph	PIPA-relation; PISC	N	social relationship recognition	object-object interaction; atmosphere	*friends, couple, professional relationship*	social signal processing

Year	Work	Image Type	Dataset	Own Data	Task(s)	Related Semantic Unit	Semantic Unit Examples	CV Cluster(s)
2020	223	natural photograph	PIPA-relation; PISC	N	social relationship recognition	object-object interaction	*colleagues, lovers, audience*	social signal processing
2020	280	natural photograph	SWiG	Y	grounded situation recognition	actions; implied actions; object purposes; object relationships	*obstacle, victim, destination*	situation recognition
2021	2	artwork images	ArtEmis	Y	affective image captioning; dominant emotion prediction	feelings; abstract concepts	*awe, freedom, love*	visual sentiment analysis
2021	154	image advertisements	Ads Dataset	N	persuasive atypicality detection	object purposes; object-object interactions	*atypicality*	visual rhetorical analysis; situational analysis
2021	183	natural photographs of everyday scenes	Intentonomy	Y	intent recognition	object purposes; purpose/meaning	*harmony, mastery, perseverance*	visual rhetorical analysis
2021	355	images from news sources	Politics	Y	political bias prediction	symbolical values; intrinsic meaning	*diversity, tradition, homelessness*	visual rhetorical analysis
2021	358	natural photographs of faces	AffectNet, SEWA, AFEW-VA	N	categorical emotion recognition; valence estimation; arousal estimation	emotion	*anger, calming, depressed*	visual sentiment analysis; social signal processing

Table I.2.2: Overview of surveyed research articles from top CV venues, focusing on highlevel semantic analysis tasks in still images. The table includes information on the year of publication, the work's title, image type, dataset used, whether the work used its own data, the specific tasks addressed, related semantic units, examples of semantic units, and the CV clusters associated with these tasks. The table offers a comprehensive look at various highlevel semantic analysis tasks, ranging from aesthetic quality inference to political bias prediction, and the semantic units and clusters involved in these tasks.

I.2.4.1 Clustering High-Level CV Tasks

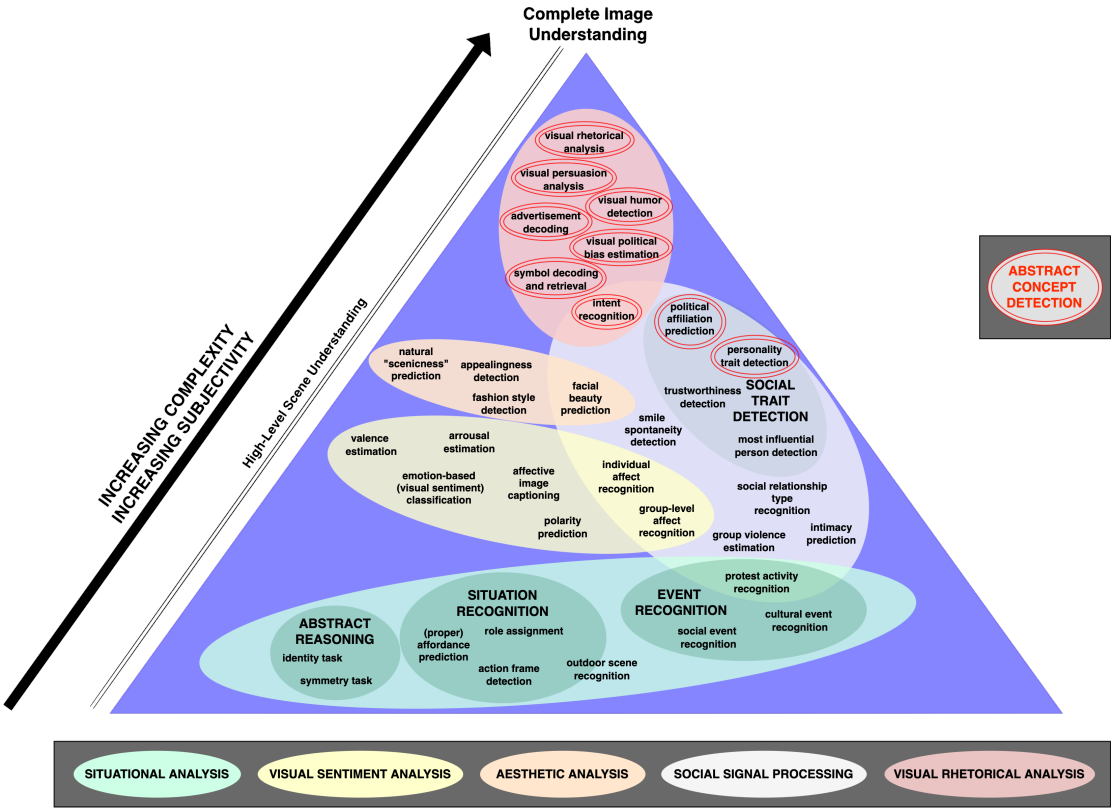


Figure I.2.3: Computer Vision tasks that deal with "high-level semantics" or "high-level visual understanding", which have been mapped also to the previous multi-disciplinary characterization of high-level semantics. Circled in red are the tasks that were found to implicitly or explicitly deal with AC detection.

Based on these findings, we formulated a set of clusters for CV tasks associated with high-level semantic units, and we assigned each publication to one or more of these clusters (as shown in the last column of Table I.2.4). Subsequently, we leveraged these outcomes to construct a cluster-based diagram illustrating CV tasks linked to high-level visual reasoning (see Fig. I.2.3. This diagram partly mirrors the structure of our high-level visual semantic unit diagram (Fig. I.2.2). In this way, the diagram serves as an alignment of the extensive body of work in CV pertaining to high-level visual reasoning and visual semantics with the same conceptual and terminological framework expounded upon in Section I.2.2. We identified five main clusters of related CV tasks that deal with high-level semantic units:

- **Situational analysis.** This cluster aligns with the commonsense semantics category, focusing on actions, activities, roles, and object purposes, along with the logical or common interactions among these elements. Within situational analysis, we can discern three primary sub-clusters of CV tasks. Firstly, abstract reasoning tackles global semantic tasks rooted in logic, such as assessing the symmetry of a chessboard or identifying its identity in various images, as exemplified in [337]. The second sub-cluster, situation recognition, as initially articulated by [389], revolves around summarizing the content of an image comprehensively. This entails identifying the primary activity, involved actors, objects, substances, locations, and most notably, the roles these participants assume within the activity. This structured prediction task goes beyond merely predicting the most prominent action, aiming to forecast the verb and its frame, which consists of multiple role-noun pairs. Finally, a substantial body of work is dedicated to event recognition, with a specific focus on social and cultural events.

- **Visual sentiment analysis.** This cluster aligns with the previously identified emotional semantics category and is also referred to as Image Emotion Analysis (IEA). Its primary goal is to comprehend how images elicit emotional responses in individuals. Despite being relatively recent, this subdomain has witnessed substantial growth in recent years, resulting in an extensive body of research and several surveys [276, 405]. Most studies have concentrated on emotion detection, aiming to identify emotions like fear, sadness, excitement, and contentment within natural images (as surveyed by [405]). Some research has delved into the analysis of group emotions in images, as outlined in [368]. In addition to natural images, this field has explored other visual media types. Notably, automatic emotion detection has been explored in the context of art images [69, 2] and memes [326].

- **Aesthetic analysis.** This cluster aligns with the previously identified aesthetic semantics category. Although relatively limited in volume, research within this cluster predominantly revolves around the detection or prediction of aesthetic value in images. These investigations encompass a range of image types, including natural scenes [385], images of human faces [149], and images as a whole [98].

- **Social signal processing.** This distinct cluster does not directly align with any of the previously identified clusters of the high-level semantics pyramid. It encompasses research related to Social Signal Processing, a broad field dedicated to constructing computational models for sensing and understanding various human social signals, including emotions, attitudes, personalities, skills, roles, and other forms of human communication. Within this cluster,

several endeavors utilize CV techniques to detect non-concrete social aspects within images. In the realm of natural images, efforts have delved into discerning persuasive intent in political photographs [187], identifying deception [57, 396], categorizing social relationship types (e.g., kinship, friendship, romantic, and professional relationships) [221], and gauging intimacy [79]. Group or crowd images have garnered particular attention, with various methodologies developed to automatically recognize non-concrete group attributes, including social cohesion [173, 22], leadership [336], warmth, and dominance [402], excitement [365], and engagement in student groups [364]. However, social signal processing tasks in the context of art images remain relatively underexplored, with some research focusing on automatically detecting the trustworthiness of depicted individuals in art images [308].

- **Visual rhetorical analysis.** This cluster exhibits a strong correlation with inductive interpretative semantics, focusing on understanding the intrinsic and implicit meaning and the purpose of images, akin to the concept of inductive interpretative semantics. It encompasses research related to visual persuasion and rhetorical techniques. Notably, much of this work has been concentrated in the subdomain of automatic advertisement understanding. While the visual rhetoric of images has traditionally been explored in Media Studies, the computational study of this field has gained prominence in the last five years, with a significant contribution from the authors of [174]. Within this cluster, one notable sub-cluster revolves around visual persuasion [187]. This subfield assesses whether images of politicians present them in a positive or negative light by analyzing facial expressions, gestures, and image backgrounds, using features to discern the persuasive intent behind the visuals.

It is worth noting that categorizing certain works into distinct clusters can be challenging, as some research efforts overlap between these domains. For instance, [143] explores the impact of group-level affect in identifying the most influential person within images of groups. This study straddles the subdomains of both visual sentiment analysis and social signal processing, highlighting the interconnected nature of these research areas.

I.2.4.2 Discussion on High-Level CV Tasks

Social and Sociocultural Emphasis in High-Level Visual Semantics

The survey encompasses a wide array of semantic units (see Table I.2.4), but social aspects, including emotions, relationships, and social events, constitute one of the most recurrent themes. This emphasis on human-centered concepts underscores

Semantic Unit Type	Examples	References
events	dining, roadtrip, parade	[127, 56, 395]
relationship type	siblings, colleagues, lovers	[374, 220, 346, 147, 223]
relationship intimacy	friendliness, warmth	[403]
social dimension	leader, influential person	[143]
personality trait	dominance, competence, neuroticism	[244, 186, 172, 321]
communicative intent	trustworthy, powerful, mastery, harmony	[172, 183]
occupation	waiter, clergy	[324]
political affiliation	democrat, republican	[187]
political bias	liberal, conservative	[355]
aesthetics	female beauty, aesthetic value	[98, 149]
atmosphere	crime rate, scenicness	[198, 385]
fashion style	hipster, bohemian	[199, 171]
humor	humor, funninness	[72]
object purposes	obstacle, audience, destination, affordance	[389, 222, 81, 344, 280]
object interaction	identity, symmetry	[337]
symbolical values	danger, comfort, freedom	[174, 390, 355, 391]
rhetorics	atypicality	[154]

Table I.2.3: Types of High-Level Semantics Extracted from Images

Image Type	Datasets	References
natural images	ImageNet [105], Occupation Database [324], PEC Data Set [56], Situnet a.k.a imSitu [389], PyschoFlickr [89], PISC [221], PIPA-relation [346], ScenicOrNot [385], ADE-Affordance [80], SWiG [280], Intentonomy [183], WIDER [395]	[98, 324, 56, 198, 389, 321, 221, 346, 385, 222, 81, 344, 147, 223, 280, 183, 395]
facial images	AffectNet [259], SEWA [204], Social Relation Dataset [403]	[186, 149, 244, 401, 358]
fashion images	Style Embedding Dataset [171]	[199, 171]
group images	Images of Groups [127], GAF-personage [143]	[127, 374, 143]
political images	Visual Persuasion [187], Politics [355]	[186, 187, 172, 355]
advertisements	Ads Dataset [174]	[174, 155, 390, 391]
artworks	ArtEmis [2]	[2]
clipart	AVH Dataset [72]	[72]

Table I.2.4: Types of Images Analyzed for High-Level Semantics

the growing significance of sociocultural elements in high-level visual semantics research, and highlights a substantial and growing body of research in CV that focuses on social aspects like social relationship recognition, visual rhetorical analysis, and situational analysis. This signifies a notable shift towards understanding the intricate interplay between visual content and societal contexts, reflecting the increasing importance of sociocultural elements in this research domain.

Diversifying Image Types

While the majority of surveyed works (approximately 75%) predominantly revolve around natural photographs, shedding light on the importance of real-world imagery in high-level semantic research, a notable inclusion of publications dealing with cultural images stands out. These cultural images span artistic, historical, and advertisement domains, hinting at an evolving inter- or cross-disciplinary approach within these types of CV tasks. This evolving trend underscores the growing relevance of cultural images and signifies that CV is extending its reach beyond conventional photography, engaging with diverse and culturally significant visual contexts.

Task-Specific Dataset Creation

Particularly noteworthy is the finding that an overwhelming majority, approximately 75%, of studies exploring high-level semantics have embarked on the creation of bespoke datasets. This observation hints at the inherent complexities in attempting to generalize across a wide spectrum of cognitively intricate tasks, where the one-size-fits-all approach of general-purpose datasets may fall short. This trend reflects the intricate and task-specific nature of endeavors related to high-level semantic units, often requiring datasets customized to address the nuances of the research objectives.

Research Output and Transformative Moments

The survey results demonstrate a pronounced rise in publications concerning high-level semantic units in top CV venues over a 15-year duration. Substantial increases in these types of publications in top CV venues seem to coincide with two pivotal moments for the CV field, 2012 and 2017 (see Figure I.2.4). Specifically, our results suggest a potential correlation between higher interest in high-level tasks in CV research and the rise of DL around 2012, impulsed by the introduction of AlexNet [211]. Before this breakthrough, CV predominantly relied on traditional ML techniques and manual feature engineering, facing difficulties in high-level semantic tasks. However, the advent of deep neural networks revolutionized the field by enabling autonomous learning of hierarchical features from data, and Convolutional Neural Networks (CNNs) played a pivotal role in these developments, empowering the field to tackle intricate aspects of visual semantics. Our survey specifically underscores a substantial increase in publications related to high-level visual understanding post-2012 which coincides with the greater trend in the field of CV. Building on this momentum, CV entered another transformative phase in 2017, marked by significant advancements in various subfields. This period expanded the scope of high-level semantic tasks, driven by innovations

such as YOLO (You Only Look Once) and its improvements [292, 291], which revolutionized real-time object detection, Mask R-CNN [164], which extended the popular Faster R-CNN framework to enable instance segmentation to identify pixel-level object boundaries, and progress in pose estimation [194], among others. The strong increase in publications after 2017 suggests that researchers in CV recognized that they now possessed the tools and methodologies to delve deeper into complex abstract semantics within visual data, propelling the field to tackle previously challenging tasks.

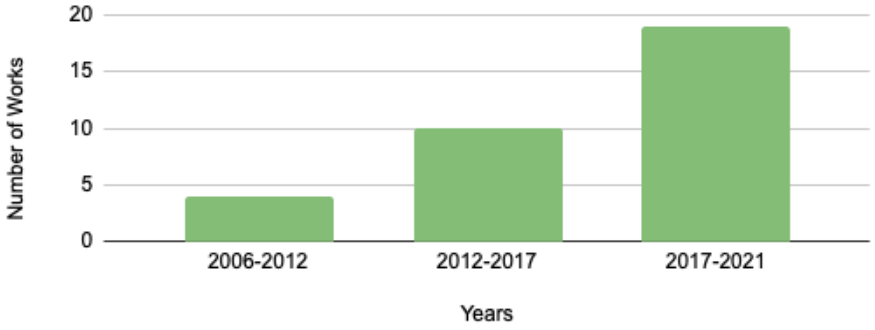


Figure I.2.4: Two inflection points, (2012) and (2017), that seem to correlate with the increasing interest in CV tasks dealing with high-level visual semantics.

I.2.5 In-Depth Survey of ACs in CV

As explained in Section I.2.3, to align our research focus with the association of ACs to visual data, we initiated a systematic subselection process to identify the works most closely dealing with tasks analogous to AC image classification. We specifically sought works emphasizing ACs as socially shared meanings embodying values and ideologies, so we did a targeted approach to study works addressing semantic units termed "symbols," "intent," and "abstract topics." By narrowing our selection, we identified 8 closely related works and classified them based on various dimensions. These dimensions included explicit versus analogous/related tasks, architectural models, reproducibility, macro F1 scores, and the hybridity of the systems. The results can be found in Tables I.2.5 and I.2.6.

Overlap in Abstract Concept Examples The AC examples employed across the surveyed works exhibit striking similarity. However, it is crucial to highlight that not all works share identical target classes due to variations in vocabulary and task definition. This divergence in target vocabulary hampers the ability to

Year	Work	Title	Img Type	Dataset	Own Data	Data Size	AC Examples
2016	[8]	*Abstract Concept and Emotion Detection in Tagged Images with CNNs*	SM	NUS-WIDE	×	14K	*love, travel, beauty*
2017	[174]	*Automatic understanding of image and video advertisements*	Ad	Ads Dataset	✓	14K	*danger, fun, beauty*
2019	[390, 391]	*Interpreting the Rhetoric of Visual Advertisements*	Ad	Ads Dataset	✓	14K	*danger, fun, beauty*
2021	[354, 355]	*Predicting Visual Political Bias Using Webly Supervised Data*	P	Politics	✓	1M	*abortion, justice, democrat*
2021	[183]	*Intentonomy: A dataset and study towards human intent understanding*	SM	Intentonomy	✓	14K	*harmony, power, beauty*
2022	[191]	*Symbolic image detection using scene and knowledge graphs*	Ad	Ads Dataset	×	8K	*danger, fun, beauty*

Table I.2.5: Overview of CV studies and associated datasets for tasks closely resembling AC image classification. Ad: Advertisement, SM: Social Media, P: Political.

Year	Work	Explicit Task	Related Task	Model	Rep.	Macro F1	Hybridity
2016	[8]	AC detection	AC detection	CNN with tag-specific binary classifiers	×	0.18	Statistical
2017	[174]	Automatic ads understanding	Symbolism detection	CNN with region attention-based image classifier	✓	0.16	Statistical
2019	[390, 391]	Automatic ads understanding	Symbolism detection	Undefined classifier and knowledge base	✓	–	Hybrid
2021	[354, 355]	Visual political bias detection	Image-word alignment	CNN multimodal feature learner, visual-only test time	✓	0.25*	Statistical
2021	[183]	Intent recognition	Intent detection	CNN with perturbation approach	✓	0.23	Hybrid
2022	[191]	Symbolic image classification	Symbolism detection	Transformer with KG, CGN, and attention	✓	0.15	Hybrid

Table I.2.6: In-depth examination of model architectures and performance in CV work analogous to AC image classification. AC: Abstract Concept. CNN: Convolutional Neural Network. GCN: Graph Convolutional Network.

leverage the same datasets efficiently. There is evidently a demand for a shared dataset that compiles culturally rich images tagged with a common set of ACs. Such a resource could benefit future research by promoting data consistency and facilitating cross-work comparisons.

Diversity in Image Types While the broader category of high-level visual understanding includes a prevalence of natural images, we observe a more concen-

trated focus on images with strong socio-cultural connotations or layers in works dealing with ACs. Notably, these images predominantly encompass social media content, advertisements, and political imagery. What sets these images apart is their cultural richness, as they often convey nuanced socio-cultural information that standard natural images may lack.

Emphasis on Advertisements It is noteworthy that a significant portion of the works are centered around advertisements and employ the same dataset [174, 390, 392, 191]. This suggests that the domain of advertisements is a good domain for the task and provides a rich source of data for the exploration of ACs in images.

Creation of Domain-Specific Datasets Many of the surveyed works take the initiative to collect and present their own datasets. This underlines the critical importance of domain-specific data for the task of AC image classification.

Consistent Dataset Magnitude Most works in this survey operate with datasets ranging from 8,000 to 14,000 images. It is important to note that the dataset size remains relatively consistent across these works. However, there is an exception in the case of [354], which generates an extensive 1 million-image dataset for their primary research focus, but in a weakly supervised way. Critically, the dataset size for the subtask of image-word alignment is not explicitly reported.

Comparable Related Tasks It is evident that many of the related tasks are comparable. They often share similar target concepts, such as *beauty*. This observation indicates that there is a degree of research overlap between these tasks. This connection, although present, had not been explicitly highlighted before.

Dominance of CNN Architectures Out of the six approaches analyzed, four employ Convolutional Neural Network (CNN) architectures as the backbone of their model structures. This prevalence suggests that deep learning paradigms, particularly CNNs, constitute a cornerstone of state-of-the-art AC image classification tasks.

Significance of F1 Score and Reporting Challenges Within this body of work, it is evident that the F1 score holds greater importance as a model evaluation metric than accuracy. This underscores the essential role of the F1 score as the primary performance measure in the realm of AC image classification. However, a noteworthy challenge emerges concerning the reporting of these metrics. In specific instances, the analogous task for some works exists as a subtask within

the broader context of the work's primary objective. For example, symbolism prediction constitutes a subtask within the larger framework of understanding visual advertisements. As a result, performance scores related to the analogous tasks may not be explicitly reported for all works, as observed in [390, 391]. Notably, this reporting gap posed a particular challenge when evaluating the highest F1 score work, namely [355], where we had to compute the score ourselves.[6] This process was necessary to assess the model's performance for more abstract classes.

Performance Challenges and Critical Questions The F1 scores attained in the realm of AC image classification reveal a significant challenge: they tend to be notably lower when compared to classes of a more concrete nature. Furthermore, the work that boasts the highest F1 score [355] only marginally outperforms the second-best in this regard. This observation is intriguing, considering that the highest-performing work benefits from two distinct advantages over its counterparts. Firstly, it operates with a substantially larger dataset, comprising one million images, two orders of magnitude greater than those employed by other works. Secondly, this work leverages a training approach that might be considered considered 'cheating', as, during the training phase, the model has access to and learns from textual documents (news articles) the images were collected from. We decided to keep this work as, even though this added textual information confers an advantage, at test time the approach does not have access to text. Despite this substantial data magnitude and training advantage, the F1 scores achieved remain relatively low, illuminating the intricate and demanding nature of AC image classification tasks.

Importance of Hybrid Approaches The prevalence of hybrid models in half of the surveyed works suggests that a purely statistical approach may be insufficient for AC image classification tasks. The inclusion of symbolic knowledge, whether through intermediary features or external knowledge and reasoning, appears to be a necessity in this domain. This observation highlights the growing relevance of hybrid AI systems for addressing the complexity of AC image classification.

[6] To achieve this, we calculated the average F1 score based on the reported F1 scores for the analogous image-word alignment tasks. We excluded scores related to the names of politicians and media outlets due to their concreteness, specifically: 'cnn,' 'trump,' 'clinton,' 'donald,' 'paul,' 'fox,' 'clinton,' 'hillary,' 'obama,' and 'republican.' The resulting F1 score was calculated based on the average scores for the following words: 'administration,' 'political,' 'conference,' 'meeting,' 'prime,' 'committee,' 'host,' 'minister,' 'foreign,' 'justice,' 'bill,' 'democrats,' 'election,' 'media,' 'candidate,' 'vote,' 'speech,' 'deal,' 'Thursday,' 'voters,' 'congress,' 'abortion,' 'democratic,' 'Tuesday,' 'news,' 'racist,' 'white,' 'illegal,' 'presidential,' 'republicans,' 'supreme,' 'gay,' 'senate,' 'immigration,' and 'immigrants.'

I.2.6 Relevant Datasets

Many popular image datasets used in CV offer low or no coverage of non-concrete concepts, such as ImageNet [105] and Tencent ML-Images dataset [386]. Others, like JFT [345], an internal dataset at Google, include non-concrete classes but are not publicly released. We identified additional datasets that contain at least some images annotated with non-concrete labels.

I.2.6.1 Natural Images

- **NUS-WIDE** [80]: Comprising 27,000 Flickr images, each associated with hashtag-like tags (approximately 4,000 unique tags).
- **Open Images** [213]: Comprises 9 million images with annotations for 19,794 classes from JFT. But while some non-concrete classes like *peace*, *pollution*, and *violence* are present, they are long-tailed and explicitly designated as non-trainable.
- **MultiSense** [139]: Comprising 9,504 images annotated with English, German, and Spanish verbs.
- **VerSe** [140]: Containing 3,518 images annotated with one of 90 verbs.
- **UNSPLASH**[7]: The complete version of this dataset contains over 4.8 million high-quality photographs accompanied by 5 million keywords including some ACs.
- **BabelPic** [65]: Comprising 14,931 images tagged with 2,733 non-concrete synsets, created by cleaning the image-synset associations from the BabelNet Lexical Knowledge Base.
- **Persuasive Portraits of Politicians** [187]: Comprising 1,124 images of politicians, each labeled with ground-truth persuasive intents of 9 types and syntactical features of 15 types.
- **Politics** [355]: Comprising 1 million images tagged with left- or right-wing political bias; each image accompanied by the text (news articles) in which they were originally embedded.
- **BNID BreakingNews** [234]: Comprising approximately 10,000 images sourced from breaking news events labeled with 77 different classes. More than half of these classes are abstract representations not directly related to objects (e.g., law, policy, G20).

[7] https://github.com/unsplash/datasets. Access date: May 2021.

I.2.6.2 Art Images

- **WikiArt Emotions** [258]: Comprising 4,000 pieces of art featuring annotations for the emotions evoked in the observer.

- **ARTemis** [2]: This dataset includes 455,000 emotion attributions and explanations provided by humans for 80,000 artworks sourced from WikiArt, including visual similes, metaphors, and subjective references to personal experiences.

- **SemArt** [135]: A multi-modal dataset designed for the semantic understanding of art. It contains fine art painting images, each associated with attributes and a textual artistic comment.

- **The Tate Gallery Collection**[8]: Tate Gallery's collection metadata featuring 70,000 artworks tagged with a taxonomy covering a wide spectrum of concrete to non-concrete concepts.

- **ArtPedia** [338]: Encompasses a collection of 2,930 paintings and 28,212 textual sentences that not only describe the visual content of the paintings but also provide additional contextual information.

I.2.6.3 Advertisement Images

- **Ad Dataset** [174]: Comprising 64,832 image ads with comprehensive annotations that cover various aspects, including the topic, sentiment, persuasive strategies, and symbolic references employed in the ads.

I.2.6.4 Datasets for Connecting ACs to Cultural Images

While these datasets encompass labels or classes extending beyond traditional concrete concepts, they may not consistently offer high-quality tags for ACs. One contributing factor is the ambiguity surrounding the origins of these tags, often aggregated from online sources. In our comprehensive examination of each dataset, we have discerned select datasets where images bear explicitly recorded AC labels, denoting human-provided annotations. Informed by insights gained from scrutinizing CV works dedicated to tasks analogous to AC image classification (Section I.2.5), we have directed our attention toward cultural images, potentially featuring superior AC tags. We provide details of these datasets in Table I.2.7

[8] https://github.com/tategallery/collection. Access date: May 2021.

Dataset Name	Image Type	Size	Primary Focus
Ads Dataset	advertisements	13,938	Features ACs as 'symbolism'
Tate Gallery	visual artworks	70,000	Rich tag taxonomy that spans concrete tags and ACs
ARTemis	visual artworks	80,000	Utterances include ACs
ArtPedia	visual artworks	2,930	Visual descriptions include ACs

Table I.2.7: Datasets with explicitly recorded AC labels

Ads Dataset This dataset is one of the very few datasets that have explicit, single ACs as tags for its images. The subset with AC tags is composed of 13,938 images and prominently features symbolism associated with ACs. The dataset provides a list of 221 symbols, each accompanied by bounding boxes. These symbols often represent common abstract ideas such as *danger*, *fun*, *nature*, *beauty*, *death*, *sex*, *health*, and *adventure*.

The Tate Gallery Collection The gallery's collection metadata, consisting of 70,000 artworks, is publicly accessible through a GitHub repository. The dataset boasts a rich tag taxonomy that spans a wide spectrum of concepts. It covers both concrete (e.g., *vacuum cleaner* and *shoe*) and non-concrete (e.g., *consumerism* and *horror*) subjects under categories like "universal concepts", making it a valuable resource for exploring abstract and non-abstract concepts.

ARTemis Although this dataset primarily focuses on emotions, encompassing 455,000 emotion attributions and explanations related to 80,000 artworks from WikiArt, it extends its scope to include ACs like *freedom* and *love*. Notably, the dataset authors have conducted an analysis to gauge the degree of abstract versus concrete language within ARTemis. To measure abstractness or concreteness, they employed the lexicon introduced by [59], which assigns a rating from 1 to 5 reflecting the concreteness of around 40,000 word lemmas. In this assessment, a randomly selected word from ARTemis received a concreteness rating of 2.81, as opposed to COCO, which received a rating of 3.55 (with a statistically significant p-value).

ArtPedia Comprises 2,930 paintings and 28,212 textual sentences. Out of these sentences, 9,173 are specifically dedicated to providing visual descriptions. Upon manual examination, it becomes evident that these visual descriptions include references to ACs, similar to what can be found in ARTemis. An example of a visual description is "Mistress and servant, a *power* relationship, maybe some deeper emotional bondage," demonstrating the dataset's potential to be curated to explore and better understand ACs within visual art.

I.2.7 Discussion

In the realm of human visual sensemaking and understanding, the term '' characterizes complex, subjective, and abstract visual reasoning, albeit with a diverse focus on various semantic units. Section I.2.2 delves into multidisciplinary analysis of this "tip of the iceberg" of high-level semantics and categorizes them into clusters based on their inherent characteristics and attributes. We identified four clusters of high-level semantics. These include *commonsense semantics*, covering more objective and widely accepted semantic elements such as actions, activities, events, relationships, and object purposes, *emotional semantics*, covering aspects related to emotions, moods, emotional cues, and individualized affects; *aesthetic semantics*, which evaluates global aesthetic attributes contributing to overall image judgments; and *inductive interpretative semantics*, encompassing complex, subjective, and culturally encoded elements like ACs, symbols, and symbolical values.

Our comprehensive analysis of 38 publications from top-rated CV venues in Section I.2.4 provides a structured categorization of high-level semantic understanding in CV, facilitating a comprehensive understanding of the field's landscape. We have organized these high-level visual reasoning clusters into five main categories. *Situational analysis* focuses on actions, activities, roles, and object purposes, incorporating abstract reasoning, situation recognition, and event recognition. *Visual sentiment analysis* explores how images elicit emotional responses, encompassing emotion detection across various image types.*Aesthetic analysis* centers on detecting or predicting aesthetic value in images across different categories. *Social signal processing* delves into the recognition of non-concrete social aspects and the study of group attributes in images. Finally, *visual rhetorical analysis* concentrates on discerning intrinsic and implicit meanings in images, particularly in the context of visual persuasion and advertisement understanding. These distinct clusters collectively provide a comprehensive framework for understanding high-level visual reasoning in CV, revealing the diverse areas of research and their interconnectedness.

Several critical lessons and emerging patterns have come to light from this survey of CV high-level tasks. First, there is a notable inclusion of sociocultural elements within this research domain, with research increasingly focusing on sociocultural aspects like emotions, relationships, and social events. This shift reflects the growing recognition of the importance of societal contexts within the field. Additionally, the survey shows that researchers are diversifying their approach to image types. While natural photographs remain a significant focus, there is a clear expansion into cultural images, signaling an evolving interdisciplinary approach in CV. Furthermore, the survey highlights the prevalence of task-specific datasets, underlining the necessity of tailoring data to meet the unique demands of various research objectives. Moreover, there's a clear trajectory of research evolution in

the field. The survey demonstrates substantial growth in publications, particularly post-2012 and post-2017. This growth aligns with the ascent of deep learning and transformative moments in CV. It signifies the field's increasing awareness of its capability to address the intricate world of abstract semantics in visual data.

The in-depth survey into works dealing with tasks analogous to Abstract Concept (AC) image classification (Section I.2.5,) offers valuable insights and lessons. Notably, it becomes evident that amassing humongous amounts of data, even with datasets comprising one million images, does not necessarily guarantee high F1 scores in AC image classification. This observation challenges the prevailing notion that massive data alone can address the complexity of this task, highlighting the need for more sophisticated approaches. Furthermore, adding textual information to the training process bestows an advantage, yet the F1 scores achieved, even with this favorable setup, remain relatively low. This underscores the intricate and demanding nature of AC image classification, emphasizing the need for novel techniques beyond data augmentation.

Another important lesson from the survey is the recognition of mid-level features, such as objects and facial expressions, as potentially crucial elements in AC image classification. Understanding the significance of these intermediary features can guide the development of more effective models, improving the accuracy of classifying ACs in images. Moreover, the prevalence of hybrid approaches in the surveyed works suggests that a purely statistical approach may fall short in the realm of AC image classification. To tackle the complexities associated with ACs in images, the inclusion of symbolic knowledge, whether through intermediary features or external knowledge and reasoning, emerges as a necessity. This finding highlights the growing relevance of hybrid AI systems that can seamlessly integrate statistical and symbolic knowledge to address the multifaceted nature of AC image classification tasks.

I.2.8 Conclusions

There is a significant body of work focused on automating high-level visual understanding tasks to mirror the most complex cognitive processes and subjective sensemaking inherent to human visual perception. This survey has revealed that a noteworthy focal point within this research landscape pertains to the investigation of social and socio-cultural cues and signals. These works explore semantic units that closely align with what cognitive science classifies as "abstract concepts," encompassing social and cultural values and ideologies, denoted by various terms such as "symbols," "intents," or "abstract topics." These are intricately tied to the domain of visual rhetorics, and our comprehensive exploration of this territory has unearthed several pivotal insights with direct implications for the advance-

ment of high-level visual understanding, especially within the realm of AC image classification.

Foremost among these insights is the recognition that, even when operating with substantial datasets, including those comprising millions of images, achieving high F1 scores in AC image classification remains a formidable and persistent challenge. This observation prompts a reconsideration of the notion that accumulating vast amounts of data alone serves as the primary panacea [355]. Additionally, the incorporation of supplementary information, such as textual content, and the judicious consideration of mid-level features such as objects and facial expressions, emerges as a critical avenue for enhancing performance. Critically, the prevalence of hybrid models in AC image classification work underscores the insufficiency of exclusively relying on statistical methodologies. The imperative inclusion of symbolic knowledge, whether through intermediary features or external knowledge and reasoning, is demonstrated as an essential component in this domain. This trend accentuates the growing significance of hybrid AI systems, poised to tackle the intricate and multifaceted challenges inherent to AC image classification tasks.

Chapter I.3

Cognitive Insights into AC Representation

Summary This background chapter first outlines the types of ACs this dissertation is explicitly interested in. It then explores a central issue in contemporary cognitive neuroscience research: understanding how ACs are represented in the human brain. It discusses the two primary avenues of semantic memory research– distributional models and embodied cognition– as well as the more recent idea of the "multiple representations view," of ACs, which suggests that both distributional and embodied information coexist in the grounding problem of ACs. This approach seeks to merge sensorimotor grounding with linguistic, emotional, and social experiences, allowing more human-like semantic knowledge to emerge. The chapter also delves into the cognitive substrates of ACs, focusing on acquired embodiment, relationality, and emotionality, highlighting the complexity and multidimensionality of AC representation in the human brain. These insights serve as the foundational knowledge for the subsequent chapters in this dissertation, which leverage these cognitive concepts for practical applications in AC image classification.

I.3.1 Ontological Considerations

ACs are hard to characterize as a unitary kind because they come in great variety, including concepts as diverse as social roles, mental states, institutional, temporal, emotional, and numerical concepts [54]. Therefore, a mapping into an ontological framework is helpful. In an attempt to map these cognitive ACs to a foundational ontology, namely DOLCE [133], the initial guess is that ACs include the following ontological classes: SocialObject (most of them: concepts, descriptions, information objects, social agents, collections), Abstract (formal entities, abstract regions -including space and time), and AbstractQuality (attributes of abstracts). Specifically, the non-concrete concepts of interest in this dissertation are what are referred to as social objects in DOLCE, more specifically as social concepts in the ontological sense expounded by [245]–immaterial products whose conventional constitution involves a network of relations among social agents. Thus, the ACs intended in this work are ontologically social objects [129], that refer to "cognitive objects with a social capital" [102], that is, cognitive clusters of prototypical situations which often involve social interactions and societal dynamics. I adhere to Masolo et al.'s [245] characterization of social concepts as immaterial constructs formed within a community. These are concepts that depend on agents who, through established conventions, not only create, employ, and discuss them but also accept their significance. In the context of this work, ACs include emotions (e.g., *sadness*), as well as abstract social, cultural, economic, and political values (e.g., *freedom*, *leadership*) and ideologies (e.g., *racism*, *consumerism*).

I.3.2 The Queer Complexity of ACs

The challenges encountered in the attempts to automate the task of AC image classification underscore the intricate and dynamic nature of ACs. This is due to their ambiguity, subjectivity, and context-dependency, as, by cognitive science standards, they are "inherently queer" and transgressive concepts [247]. Unlike concrete concepts, which trigger neural areas linked to actions related to their referents, ACs defy this pattern by not consistently activating sensorimotor regions. The transgressive nature of ACs manifests in their ability to transcend the limitations of this binary confinement. Importantly, the queerness of ACs arises from their resistance to fixed, simplistic, or normative definitions, a resistance that poses both a significant technical and ethical challenge due to their elusive, subjective, and context-dependent nature. Thus, to approach the task of AC image classification more effectively, embracing this queerness and leveraging it to our advantage seems crucial. Consequently, we must gain a deeper understanding of AC representation in the human brain and explore how to harness it within the

realm of AI for associating visual data with these intricate ACs. In the field of cognitive science, "understanding how abstract concepts might be represented is a crucial problem for contemporary research" [53, p. 1]. Understanding how meaning is represented is the core problem of modern semantic memory research, which has chiefly operated on two independent paths: distributional semantic models and embodied cognition [100].

I.3.3 Distributional vs. Embodied View

Distributional models suggest that meaning can be inferred from the contexts (almost always operationalized as language contexts) in which words appear. As such, the meaning of concepts is derived from and represented in terms of statistical patterns of co-occurrence with other words in a language, based on Firth's (1957) supposition that "*you shall know a word by the company it keeps*" [122].

On the other hand, embodied approaches propose that meaning finds its roots in our sensory, perceptual, motor, interoceptive, and introspective interactions with the environment (e.g., [30]. These theories also emphasize the significance of context, as discussed in reviews by [393] and [394]. However, in embodied theories, context is considered situated and grounded. To grasp the meaning of a word, we engage in a mental simulation of the bodily experiences associated with encountering that concept "in the wild" [30]. Importantly, this simulation is context-dependent and influenced by an individual's personal history.

I.3.4 Multiple Representations View

More recently, reconciliation between these two representations of meaning has been the focus of much research, specifically by considering distributional and embodied information as fundamentally the same type of data, entangled and mutually influencing each other across multiple timescales [100]. The emergence of such "multiple representation views" has been especially crucial for ACs [14], with the latest research pointing towards their grounding in sensorimotor systems while also involving linguistic, emotional and social experiences as well as internal experiences [100] (see Figure I.3.1). In this sense, it is thought that "uniting distributional and embodied data under a common framework provides a potential solution to [...the grounding] problem of abstract concepts'" [100, p. 5], with attempts being made to conjunctly represent all these different types of information. Increasingly, emergent representations are not simply the sum-total of feature-based and distributional linguistic representations: it is the interaction between experiential and linguistic data that allows for more human-like semantic knowl-

edge to emerge [15]. In this sense, we *know words by the linguistic and perceptual company they keep* [232].

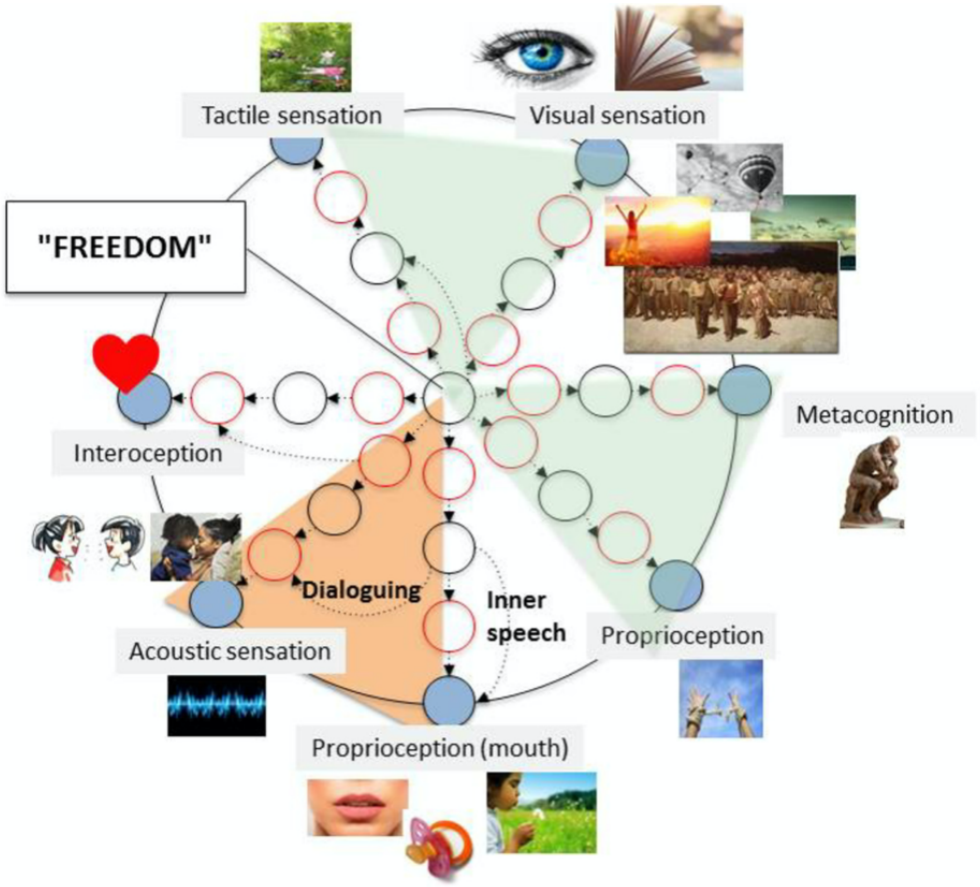


Figure I.3.1: The multidimensional grounding of the AC "freedom" according to cognitive science WAT theory. As explained by [52] Similarly to concrete concepts, *freedom* has the potential to activate sensory modalities, as well as interoception and proprioception. For instance, when it evokes scenes like lying on the grass with friends and gazing at the sky, it may elicit the tactile sensation of the cool grass against the body. Alternatively, when simulating the act of freeing oneself from a rope, it might reproduce proprioceptive sensations associated with the body's movements while constrained by the rope. This multifaceted grounding accounts for the diverse array of visual sensations associated with the AC of "freedom." The figure is a direct replication from [54].

I.3.5 Cognitive Substrates of Abstract Concepts

In cognitive science research, key features and representational substrates of ACs are the focus of much research [369, 141, 167]. Three cognitive aspects of ACs are of special interest to this work, as they are potentially translatable into computational frameworks: acquired embodiment, relationality, and emotionality.

I.3.5.1 Acquired Embodiment

The multiple representations view has been modeled in neuroscientific research that merges sensory-motor (S-M) information with distributional statistics from natural language [15, 111, 340], demonstrating how S-M knowledge linked with a particular word can be indirectly extended to its lexical associates [185]. Hoffman and McClelland (2018) [167] introduced a computational framework known as the 'hub-and-spoke model', which combines a hub-and-spoke architecture with a buffer that can be influenced by prior context [167]. Building on their findings, they propose the concept of *acquired embodiment* to explain how abstract words become connected to sensory-motor (S-M) information through their associations with concrete words. In their research, as shown in Figure I.3.2, they present evidence illustrating this concept. They plot the activations for S-M features shared by all members of a category when the network encounters both representative concrete and abstract words. For concrete words, the network is trained to activate the corresponding S-M features upon encountering them, resulting in a clear, binary pattern of S-M activation. In contrast, during training, abstract words do not provide targets for S-M units, aligning with the notion that ACs lack direct associations with sensory-motor experiences. Consequently, the activity of these units remains unconstrained during the learning process. Nevertheless, as demonstrated in the figure, when presented with abstract words, the network gradually activates S-M features associated with the concrete items with which they regularly co-occur.

An important distinction in Hoffman et al.'s [167] model of acquired embodiment is its context-dependent nature. This is exemplified in the bottom half of Figure I.3.2, where distinct S-M activations are observed for the same abstract words in varying contexts. For instance, when *journey* follows *cashier*, it triggers strong activation of the vehicle-related S-M units due to the frequent co-occurrence of the two words in discussions about modes of transport. Conversely, when *journey* follows *duchess*, it induces weak activation, as the two words rarely co-occur in contexts related to vehicles. Therefore, the specific S-M information activated by abstract words hinges on the specific context in which they are encountered, aligning with findings that demonstrate how context shapes the types of sensory-motor knowledge participants access in response to words [167].

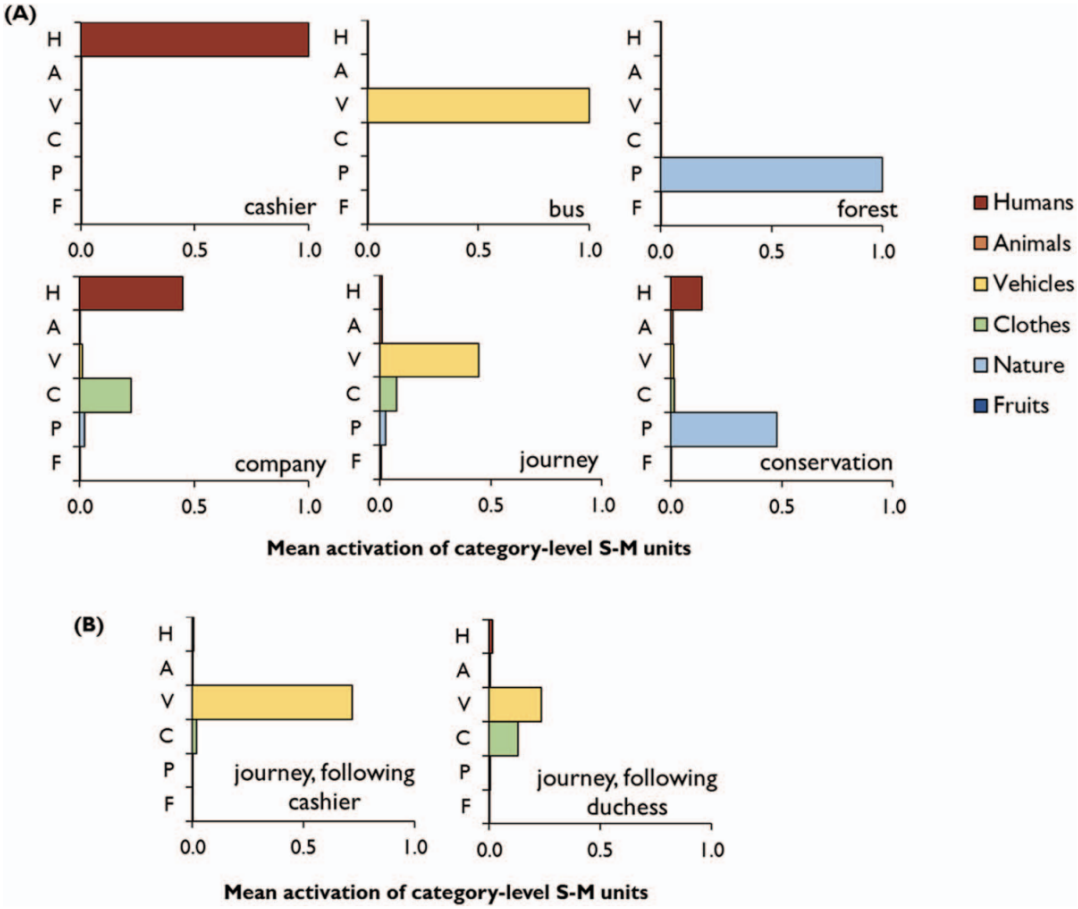


Figure I.3.2: Abstract words come to be linked to sensory-motor (S-M) information by their associations with concrete words—a process called "acquired embodiment." S-M unit activations for a selection of concrete and abstract words. (A) Activations of S-M units shared by the members of each category, in response to a selection of words. (B) Activation of S-M units in response to the same abstract word in two different contexts. Figure reproduced directly from [167].

The concept of acquired embodiment presents an intriguing avenue for enhancing computer vision tasks, as it holds potential for improving the representation of images with regard to ACs in machine-readable formats. A notable advantage of this concept lies in its capacity to deduce experiential attributes for ACs that might otherwise lack robust sensorimotor associations or possess fewer sensory-perceptual links. For instance, if the AC *death* frequently co-occurs with the term *coffin*, which strongly embodies sensory-perceptual attributes such as the color *black*, we can consider utilizing this acquired association between *black* and *death*

to influence image representations or guide the training process of image classification based on these complex semantic notions.

I.3.5.2 Relationality

According to [123], ACs are content-flexible schemata: specifiers of multiple semantic relations between entities, which allow for a fairly unconstrained range of entities to fill the schema. For example, the concept of *difference* can emerge from the relationship(s) between two objects compared on a physical dimension, such as their shape, as well as from the relationship(s) between two people, compared on a mental dimension, such as their emotional responses to an event. Advances in cognitive science posit that, as opposed to concrete concepts, ACs rely on semantic rather than categorical similarity relations [91] and on associative relations [110]. In other words, ACs are specifiers of relations between entities, allowing for a fairly unconstrained range of fillers to fit the schema. As such, a possibility is that the detection of ACs may rely on the detection of other types of semantic relationships between discrete objects in images. However, detecting such relationships and interactions tends to be a cognitive task, integrating perceptual information into conclusions about the relationships between objects in a scene [58, 121]. A cognitive understanding of our visual world thus requires that we complement computers' ability to detect objects with abilities to describe those objects [180] and understand their relationships within a scene [306].

I.3.5.3 Emotionality

The idea that internal, and especially affective, states may play a role in the representation of abstract words and concepts is not new [12, 11, 29]. [205] demonstrate that emotional content plays a crucial role in the processing and representation of ACs, showing that statistically, abstract words are more emotionally valenced than concrete words. They propose that the acquisition of words denoting emotions, moods, or feelings may be a crucial stepping stone in the development of abstract semantic representations, and that differences between concrete and abstract words arise because of a general statistical preponderance of affective information for abstract words (and sensorimotor information for concrete words). [369] also report a functional magnetic resonance imaging experiment that shows greater engagement of the rostral anterior cingulate cortex, an area associated with emotion processing in abstract processing. A correlation analysis of more than 1,400 English words further showed that abstract words, in general, receive higher ratings for affective associations (both valence and arousal) than concrete words, supporting the view that engagement in emotional processing is generally required for processing abstract words.

Part II

Defining + Benchmarking AC Image Classification

Chapter II.1

The ARTstract Image Dataset: AC Visual Evocation

Summary In this chapter, we introduce the ARTstract image dataset, a valuable resource for investigating the intersection of visual data and ACs within the realm of CV and computational visual studies. ARTstract is curated from four diverse datasets, combining high-resolution cultural images related to ACs. This dataset addresses challenges related to AC labeling in cultural images, and provides a tool for researchers to explore the interplay between visual content and conceptual meaning. The chapter discusses the dataset's creation, integration, composition, and statistics, revealing a significant class imbalance among AC clusters. Additionally, the chapter touches on the challenges of defining and capturing ACs within the dataset, acknowledging their cultural and contextual nuances and biases. Despite these limitations, ARTstract serves as a foundation for further research in areas such as digital humanities, art history, and cognitive science, offering a valuable resource to experiment with explainable computer vision methods and inspire the development of more culturally sensitive and diverse AC image datasets.

II.1.1 A Novel Resource for Investigating ACs

In alignment with our exploration of the intricate challenges posed by ACs within the realm of computer vision and computational visual studies, we introduce the ARTstract dataset, a resource that addresses the gaps and complexities discussed thus far. This dataset presents a tool for researchers seeking to delve into the intersection of visual data and abstract ideas. Comprising an array of high-resolution images encompassing cultural artworks and advertisements, each associated with ACs, ARTstract emerges as a significant asset to advance our understanding of the interplay between visual content and conceptual meaning.

As discussed in Section I.2.6 of Chapter I.2, while many image datasets offer labels for concrete concepts, it can be challenging to find high-quality tags for ACs due to the ambiguity in their origins, often collected from online sources. One significant insight from our survey was the identification of datasets that explicitly record AC labels, potentially providing superior data for training models for the task of AC image classification. Specifically, four datasets were identified as potential resources that could be combined, as they all share a common focus on cultural images and include ACs in labels or metadata.

As such, ARTstract was curated by combining data from these four datasets, namely ArtPedia [338], ARTemis [2], the Ads Dataset [174], and the Tate Collection metadata. The inclusion criteria for images were twofold: relevance to the task of AC image classification and high resolution. By adopting this approach, ARTstract encapsulates a diverse range of visual materials that have inherently interwoven abstract ideas within their imagery.

II.1.2 Data Sources

The Ads Dataset (ADVISE) [174, 390] This dataset includes over 64,000 image ads covering a diverse range of subjects. Each image ad is tagged with its topic, the sentiment it attempts to inspire in the viewer, and the strategy it uses to convey its message. The dataset also includes "symbols" that the ads use, a common technique used in advertising to convey meaning and emotions to the viewer, such as the concept of "peace" symbolically represented by a dove. The dataset includes 13,938 ad images with 221 AC symbol tags, each with corresponding bounding boxes, and then clustered into 53 symbol clusters. The most common symbolic ACs are *danger*, *fun*, *nature*, *beauty*, *death*, *sex*, *health*, and *adventure*.

The Tate Gallery It houses the United Kingdom's national collection of British art, as well as international modern and contemporary art. Their collection meta-

Figure II.1.1: ARTstract contains cultural images including paintings, visual artworks, and advertisements, such as the ones pictured evoking the AC of "death". Clockwise from top left: *The Apotheosis of War* (1871) by Vasily Vereshchagin, Tretyakov Gallery, Moscow, Russia, public domain; *Panasonic: Where no vacuum has gone before* (2014) advertisement by Y and R; *Christ Carrying the Cross* (1660) by Jacob Jordaens, Rijksmuseum, Amsterdam, Netherlands, public domain; *The Axe Effect* (2003) advertisement by Lowe Bull Calvert Pace; *The head of Christ* (1864) by Edouard Manet, public domain; *Mess of fish* (1940) by Paul Klee, public domain.

data of 70,000 artworks, available as a Github repository,[1] includes complete records of most artists and artworks in the collection, along with image and thumbnail URLs. The Tate's subject taxonomy for labeling their artworks is a unique feature of the dataset, including a wide spectrum of subject tags. The taxonomy was expert-led, developed alongside the digitization of the Tate's collection, and organized in a hierarchical structure. Critically, the rich tag taxonomy covers both concrete (e.g., *vacuum cleaner* and *shoe*) and non-concrete (e.g., *consumerism* and *horror*) subjects under categories like "universal concepts", making it a valuable resource for exploring abstract and non-abstract concepts.

ARTemis [2] A large-scale dataset providing data about the interplay between visual content, its emotional effect, and its language explanations, ARTemis focuses

[1] https://github.com/tategallery/collection. Access date: March 2021.

on the affective experience triggered by visual artworks and asks annotators to indicate the dominant emotion they feel for a given image, as well as to provide a grounded verbal explanation for their emotion choice. The dataset contains 455K emotion attributions and explanations from humans on 80K artworks from WikiArt. It provides a rich set of signals for both the objective content and the affective impact of an image, creating associations with ACs, such as *freedom* and *love*. Notably, the dataset authors have conducted an analysis to gauge the degree of abstract versus concrete language within ARTemis. To measure abstractness or concreteness, they employed the lexicon introduced by [59], which assigns a rating from 1 to 5 reflecting the concreteness of around 40,000 word lemmas. In this assessment, a randomly selected word from ARTemis received a concreteness rating of 2.81, as opposed to COCO, which received a rating of 3.55 (with a statistically significant p-value).

ArtPedia [338] A dataset of paintings with both visual and contextual descriptions, it contains over 2,930 images, and the manual annotation of each sentence as either visual or contextual allows for a comprehensive analysis of the visual and semantic content of the dataset. Additionally, ArtPedia is the only dataset to contain both types of artistic sentences, making it a valuable resource for developing visual-semantic models capable of jointly discriminating between visual and contextual sentences of the same painting. Upon manual examination, it becomes evident that these visual descriptions include references to ACs, similar to what can be found in ARTemis. For instance, some examples from the dataset illustrate this well: "Mistress and servant, a *power* relationship, maybe some deeper emotional bondage," or "The foam might suggest that the tree is caught on an unseen rock; there is ambiguity in whether this location is a small respite of stability or highlights the imminent *danger* of reaching the fall's edge."

II.1.3 AC Selection and Definition

The ARTstract dataset employs evoked clusters as a method for labeling the ACs represented in each image. Evoked clusters are collections of ACs that frequently appear together in specific contexts. This concept of clustering abstract symbols for their visual evocation was first introduced in the Ads Dataset [174] and has been adopted in limited prior research on AC image classification within the realm of computer vision, as seen in works like [390] and [191]. While it is critical to note that these cluster categories are not flawless and definitely lack complete objectivity, they are currently the benchmark for AC image classification. This is why we have chosen to retain them in the creation of this dataset. The original clusters were derived from an analysis of the co-occurrence of ACs in advertising

images, and some choices made in the cluster creation process reflect a Western bias, such as the inclusion of the term "america" within the *freedom* cluster, which does not universally apply. However, we have preserved the clusters as they are for two key reasons. First, we aimed to ensure that our work could be compared to existing research, thus advancing the state of the art. Second, our focus is on understanding the biases embedded within these concepts rather than their universality. Nevertheless, we strongly advocate for future research to explore more culturally-sensitive and diverse AC datasets.

Beginning with the clusters sourced from the Ads Dataset, we aimed to curate a subset of clusters that held the greatest relevance for our study on ACs from a cognitive science perspective. To achieve this, we adopted an approach grounded in cognitive science research insights. Specifically, we cross-referenced the clusters from [174] with a comprehensive list of ACs derived from foundational cognitive science work [160]. In this process, we identified clusters containing words that met two specific criteria: a) they were present in the cognitive science study of [160], and b) they garnered abstractness ratings under 3.75 in (the lower the more abstract), following cognitive science standards as presented in [160]. By leveraging the findings of cognitive science research, our goal was to pinpoint a diverse range of ACs that would significantly contribute to our exploration of the intricate relationship between ACs and visual imagery. To further refine our selection, we deliberately excluded AC clusters that had already been extensively examined in the computer vision literature, such as emotions (e.g., *happiness* and *love* [257]) and *violence* [290]. After this meticulous selection process, we identified seven clusters of ACs, with the cluster words defined by [174], which we found to be well-suited for our research. These clusters are:

- comfort: *comfort, cozy, soft, softness*
- danger: *danger, peril, risk*
- death: *death, lethal, suicide, funeral*
- fitness: *exercise, fitness, running*
- freedom: *america, freedom, liberty*
- power: *force, power, powerful*
- safety: *safety, security*

II.1.4 Image Mining and Processing

We aligned images from the four datasets with the chosen clusters. This matching process involved distinct procedures for each of the original data sources. In the

cases of the Tate and ADVISE datasets, individual images have words as individual tags. For each AC cluster, we systematically pinpointed images bearing at least one of the cluster's associated words, and simultaneously recorded the frequency of usage of these cluster words, indicative of their evocation strength. As for ARTemis and ArtPedia, we delved into the "'utterances" and "visual sentences," respectively, seeking instances where these contained any of the cluster-evoking terms. Upon discovering such instances, we deemed them as evocations. In parallel with our approach in the other datasets, we tallied these evocations, and detailed and comprehensive information about these steps can be accessed within the GitHub repository.[2] An example of the kinds of images that were collected is visible in Figure II.1.2.

Figure II.1.2: Cultural images tagged with the AC *danger* in ARTstract. This example shows that AC labels are semantically diffused and associated with visually variant images. From left to right: *The Roaring Forties* (1908) by Frederick Judd Waugh, Metropolitan Museum of Art, public domain; *The Hippopotamus and Crocodile Hunt* (1615) by Peter Paul Rubens, Alte Pinakothek in Munich, Germany, public domain; Untitled advertisement by Telecinco against domestic violence; *Tales of Mystery and Imagination by Edgar Allan Poe* (1923) by Harry Clarke, Metropolitan Museum of Art, public domain.

II.1.5 Dataset Integration and Composition

In the ARTstract dataset, each of the 14,795 images is associated with a single AC cluster. The images are in JPG format with a resolution of 512x512 pixels. For each image assignment, we diligently tracked various metadata, including the source dataset and unique identifier (ID), alongside crucial details such as the evocation strength and evidence of the evocation. This comprehensive documentation process allows for in-depth analysis of the dataset's composition and insights into

[2] https://github.com/delfimpandiani/ARTstract_Seeing_abstract_concepts. Last access date: February 2024.

Split	Total	Comfort	Danger	Death	Fitness	Freedom	Power	Safety
Train	11818	4984	1124	2207	632	409	2174	288
Val	1485	614	138	280	92	46	280	35
Test	1492	603	170	257	102	51	276	33
Total	**14795**	**6201**	**1432**	**2744**	**826**	**506**	**2730**	**356**

Table II.1.1: ARTstract Dataset Statistics

the association between visual imagery and ACs. The ARTstract dataset boasts a total of 14,795 images, with each image being labeled with one of the seven AC clusters. These images are provided in JPG format and maintain a consistent resolution of 512x512 pixels.

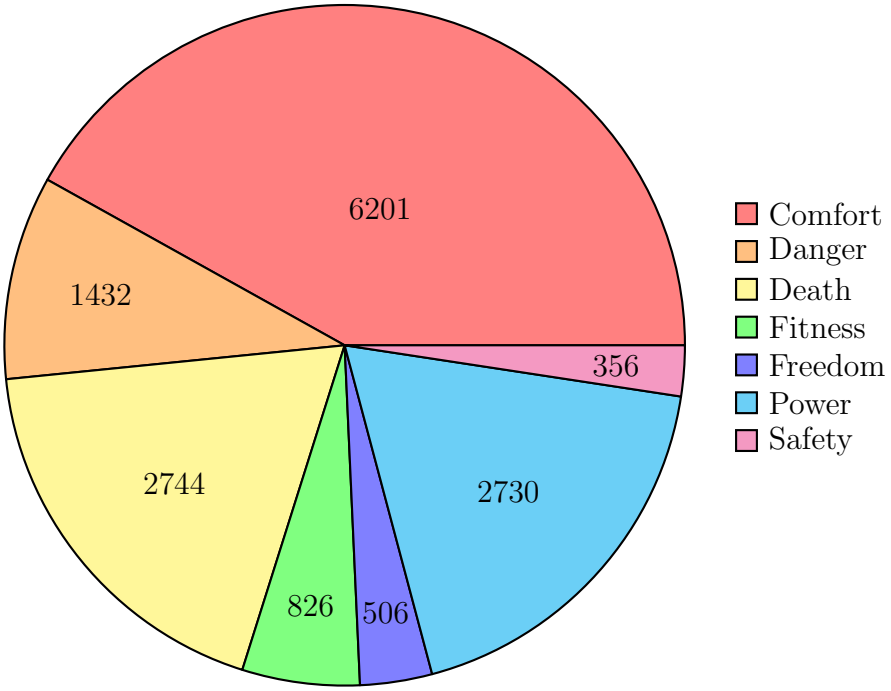


Figure II.1.3: Pie Chart of Concept Distribution

For detailed statistics and a breakdown of the dataset by cluster assignments, please consult Table II.1.1. This table offers a comprehensive view of the dataset, showcasing the total number of images and their distribution across the seven distinct AC clusters in the training, validation, and test splits, providing valuable insights into the dataset's composition. In addition, Figure II.1.3, depicted as a pie chart, highlights the distribution of AC clusters in the ARTstract dataset. Notably, the dataset displays a significant class imbalance, with certain clusters, including *comfort*, *death*, and *power* contributing substantially to the dataset. This

visualization offers an initial overview of the dataset's composition. To mitigate the class imbalance's potential impact on model generalization, we also offer a "balanced" subset of ARTstract. In this subset, clusters with over 1,000 instances are capped at 1,000 images, while clusters with fewer than 1,000 instances retain all their images, ensuring a more balanced cluster representation. See Figure II.1.4 for the statistics of the training, validation, and test datasets.

II.1.6 ARTstract and its Coverage

The ARTstract dataset is a new resource for cultural heritage and computer vision research, particularly useful for studies pertaining to digital humanities, art history, and cognitive science. However, some key limitations of the dataset are related to the inherent nature of attempting to create a clear definition of an AC. Furthermore, as we chose to reuse and combine existing datasets, ARTstract is bound by choices made by the creators of these datasets. We acknowledge that ACs are inherently culturally motivated, this is perhaps most visible in the ADVISE dataset as the advertising domain is highly culturally contextualized. However, also in ArtPedia and the Tate Gallery Western (European) art makes up a larger proportion which has a direct impact on ARTstract's coverage and representation. Within these artworks, certain themes and symbols form a shared conceptual grounding to their intended audiences [278], therefore these datasets are suited to use for our purpose, with the caveat that they represent a particular context. The coverage limitation within ARTstract stems partly from the intrinsic challenge of cultural bias when assigning labels and meanings to images, particularly concerning ACs. Moving forward, it is imperative to expand coverage and enhance the explicit contextualization of labels, dispelling the notion of their objectivity.

Despite its limitations, ARTstract fills a significant gap by providing much-needed data for tasks related to AC image classification. Critically, we believe that the richness and diversity of the ARTstract dataset provide a unique opportunity for exploring and experimenting with explainable CV methods. The dataset offers a testing ground for existing and novel explainable CV methods, demonstrating the potential of combining technical methods with hermeneutic work to develop interpretable systems. Moreover, the significance of the ARTstract dataset goes beyond its value as a resource for cultural heritage and CV research. The evocation of ACs is complex, subjective, and culturally variant, and as such, we hope that the development of this dataset can be a source of inspiration to expand it with more complex, situated, and multicultural perspectives.

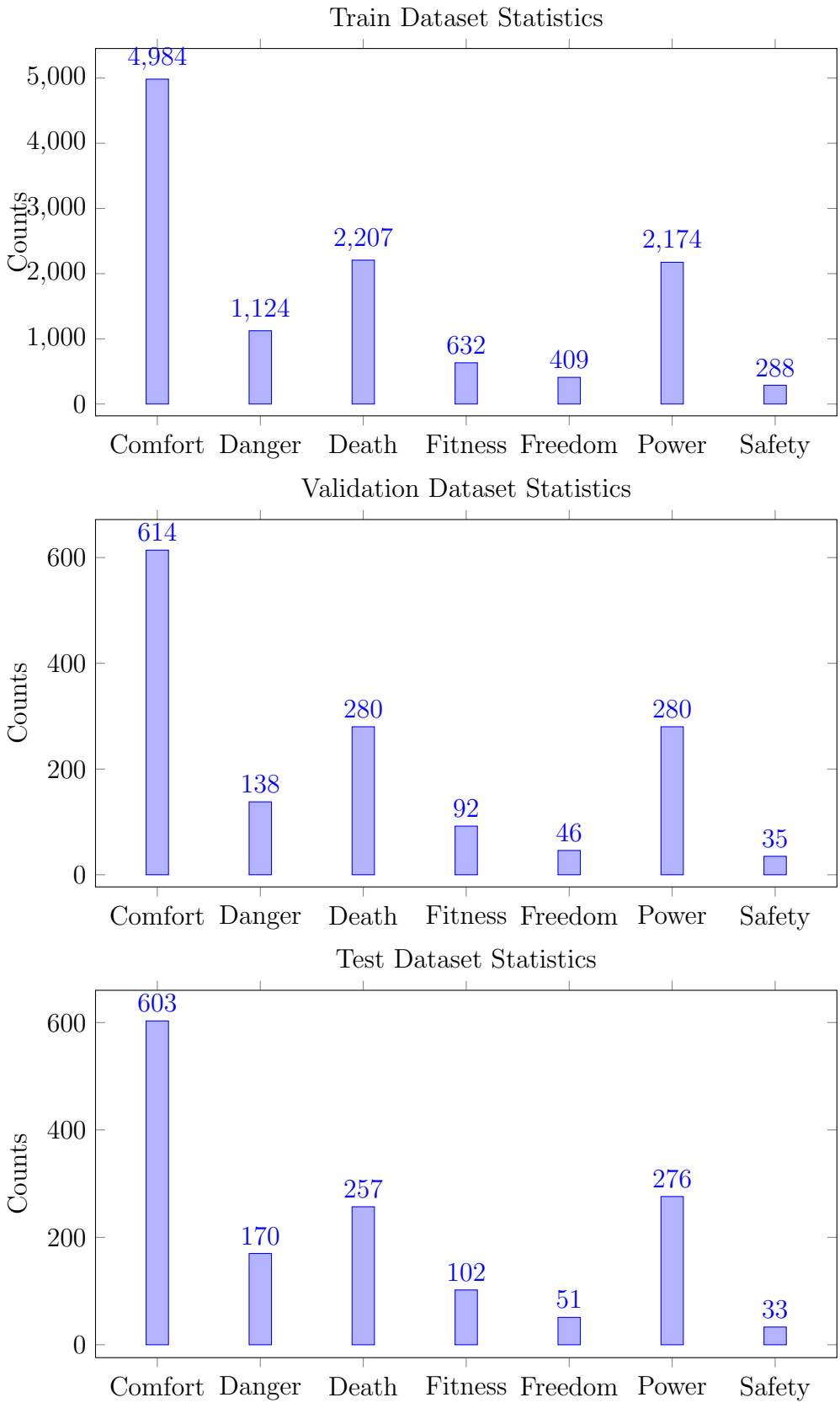


Figure II.1.4: Split-specific statistics of the distribution of ARTstract.

II.1.7 Limitations and Further Directions

While ARTstract serves as a pioneering dataset for AC image classification, it exhibits limitations regarding its diversity and size. Currently, ARTstract is predominantly Euro- and Western-centric, with a majority of its content sourced from Western projects, leading to a bias towards European and North American art and culture. Additionally, the dataset's size is relatively small compared to other art datasets such as WikiArt [258], Web Gallery of Art [70], or the TICC Printmaking Dataset [268]. This size limitation hampers the effectiveness of models that rely on large sets of input images, thereby impeding progress in state-of-the-art image classification tasks [76]. Multiple further directions can be taken with this work to start facing these limitations, including:

1. **Addressing Cultural Bias and Diversity**:
    - Identify and embed the cultural bias inherent in attributing abstract meanings to visual data through labels, in a manner that can be understood by machines [241]. Subsequent research should prioritize explicitly tracking and rendering machine-readable diverse cultural contexts and their corresponding evocations of ACs. This approach would provide alternative perspectives and insights on these concepts.
    - Enrich the ARTstract dataset with a wider range of cultural images, including those from non-Western perspectives, to capture a broader spectrum of AC evocation. This diversity would better represent the cultural richness and subjectivity inherent in high-level visual understanding.
    - Obtain human annotations of ACs from annotators with diverse cultural backgrounds to assess and enhance the reliability of ARTstract. Incorporating human-checked tags and tracking them using tools like knowledge graphs could significantly improve the dataset's utility.
2. **Dataset Size Limitations**:
    - Extend ARTstract to include more and other types of artistic creations. An interesting approach is that of extending it using automatic tagging of images and other media by using techniques from the natural language processing domain [226], such as the extraction of AC from descriptive sentences using topic modeling [197] or linguistic frames [281].
    - Expand the cluster definitions within ARTstract to encompass a wider array of words, languages, and cultural nuances. Aligning with resources like WordNet [251] and BabelNet [262] can provide a more

comprehensive definition of clusters, while relevant ontologies [26, 46] can offer semantically richer descriptions. This would allow for better integration with other relevant artwork information.

II.1.8 Conclusion

In this chapter, we introduced the ARTstract dataset, a curated collection of high-resolution cultural images associated with AC labels. ARTstract serves as a valuable resource for exploring the relationship between visual content and abstract ideas, addressing challenges related to AC labeling in cultural images. Curated from four diverse datasets, it encapsulates a wide range of visual materials with abstract ideas interwoven within their imagery. However, the dataset exhibits limitations in diversity and size, primarily being Euro- and Western-centric and relatively small compared to other art datasets. Moving forward, efforts to enhance these aspects are critical to unlocking its full potential. Additionally, mitigating cultural bias and enhancing AC labeling reliability are essential for improving the dataset's usability and effectiveness for research purposes. In conclusion, ARTstract represents a pioneering dataset for investigating ACs in visual media.

Chapter II.2

End-to-End Deep Vision: Deep Learning AC Image Classification

Summary This chapter critically assesses the efficacy of state-of-the-art deep learning (DL) models for classifying images based on ACs. Anchored by the introduction of the ARTstract dataset in Chapter II.1, this chapter delves into the utilization of feature vector representations extracted from state-of-the-art DL models, and attempts to interpret the cultural meanings assimilated by these models. The chapter is structured around three fundamental components, each addressing one sub-research question in the domain of deep AC image classification. The first section (Section II.2.4) delves into the analysis of intraclass similarity within deep representations of image clusters tagged with the same concept. It hypothesizes reduced similarity for clusters associated with ACs, due to the inherent flexibility of such concepts. The second section (Section II.2.5) is dedicated to the training and performance evaluation of state-of-the-art DL classification models, with a primary focus on Convolutional Neural Network (CNN) models for AC image classification. It anticipates potentially lower performance compared to traditional image classification tasks, and in line with F1 scores from related tasks (see Chapter I.2. The third section (Section II.2.6) delves into the realm of model explainability, recognizing the challenges in generating human-understandable explanations from black-box systems. Our approach encompasses diverse methods, including saliency map generation, perceptual topology exploration, and the introduction of a novel non-traditional explainability technique, known as SD-AM. SD-AM offers valuable insights into model decision-making and the creation of human-readable feature visualizations. In sum, this chapter significantly contributes to our understanding of AC image classification and the interpretability of DL models.

II.2.1 Introduction

The contemporary shift to an era characterized by selection and curation has elevated the importance of automatic models for navigating and comprehending vast troves of visual data. These models are now expected to handle abstract visual reasoning tasks, such as predicting personality traits [321], political bias [355], and intents [183]. As emphasized in the introduction (Chapter I.1) and survey on the state of the art (Chapter I.2) of this dissertation, a significant aspect of these evolving expectations is the central role that ACs are assuming within automated visual processing. The diversity and divergence present in visual signals evoking individual ACs, however, present a formidable challenge to conventional methodologies rooted in the CV domain. Although CNNs offer immense promise, they are intrinsically optimized for tasks characterized by high intraclass similarity [42, 328]—qualities that conflict with the inherent heterogeneity and culturally nuanced nature of ACs. Thus, the task of detecting ACs within visual data can be seen as a complex "wicked problem", lacking clear-cut solutions and molded by multifaceted cultural intricacies [297]. The increasing utilization of CNNs for wicked problem tasks like AC detection raises questions about the delicate knowledge assimilated by these models. Thus, their explainability becomes crucial in understanding how they handle complex socio-cultural visual reasoning tasks.

Explainability can be seen as a response to the perceived "explanatory deficit" in technical disciplines [44]. It stems from the challenges posed by opaque DL models and the recognition of problematic biases that can lead to inequities. The emerging subfield of eXplainable AI (XAI) advocates for interpretability as a means to address these issues [254]. In this context, the ability to explain predictions is crucial and informative of the heuristic process itself [272]. We argue that for the socio-cultural-cognitive task of AC image classification–which is based on subjective, cultural, and interoceptive processes–the perils of model reuse without explainability are high, as it can potentially echo harmful stereotypes or visions of the world based on prejudice, racism, and other biased worldviews.

This chapter critically engages with the prevalent trend of automating high-level visual reasoning via DL–placing exclusive reliance on visual signals–prominently featuring CNNs. We delve into this trend, scrutinizing the knowledge sought by CNNs and the knowledge they ultimately encapsulate. In the context of deep machine vision, we are especially interested in the explainability techniques of class activation mapping (CAM) and activation maximization (AM), also known as feature visualization (FV). CAM identifies the salient regions of an image that are considered important by the model in a classification task. It has been used, for instance, to identify where CNN models localize symbols in iconography classification [363]. FV (AM), on the other hand, involves the generation of images that visualize what a neuron in a CNN has learned [116]. It has been employed

to visualize neurons of models trained on natural images as well as on artistic images (see Figure II.2.1). Offert and Bell [272] contend that they can be viewed as technical metapictures [253] functioning as *hypericons*: indirect "illustrations," or "visualizations" in the literal sense of forcibly and subjectively summarizing and making-visual the non-visual, which can serve as summary images "a theory of knowledge" [253, p. 9] becoming crystallized in a machine learning model. Through the use of such methods, the technical system becomes an integral part of the interpretive process rather than an opaque tool.

This chapter employs the novel ARTstract dataset (see Chapter II.1) as a case study to investigate three critical aspects of the effectiveness of the DL paradigm for the task of AC image classification:

- **Deep Representation Analysis**: We examine intraclass similarity within deep representations of image clusters labeled with the same concept.
- **Performance Evaluation:** We present baseline model performances on ARTstract to benchmark image classification based on ACs.
- **Explainability Experiments**: We explore traditional and non-traditional paths for enhancing the interpretability of CNNs, inspired by [273]. To better understand how CNNs assimilate and reflect cultural meanings, and to discern the echoes reverberating within these visions, we introduce SD-AM, a novel approach to explainability. This approach condenses visuals into hypericon images through a fusion of feature visualization techniques and Stable Diffusion denoising.

Overall, this chapter critically addresses AC image classification's challenges within the DL paradigm. By embracing innovative methodologies and providing comprehensive analyses of explainability techniques, we make a substantial contribution to the broader discourse surrounding automatic high-level visual understanding, its interpretability, and the ensuing implications for comprehending culture within the digital era. Through our exploration, we illuminate the multifaceted trends, complexities, and opportunities that underlie the fusion of high-level visual reasoning and computer vision.

The remainder of this chapter is organized as follows. In Section II.2.2, we present the end-to-end deep vision paradigm that this chapter applies to the task of AC image classification. In Section II.2.3, we review related work, including computer vision work concerning image classification, explainability, and their overlaps with cultural data. In Section II.2.4 we investigate intraclass similarity within deep representations of image clusters tagged with the same concept, hypothesizing lower similarity for AC clusters compared to concrete ones due to the

Figure II.2.1: Examples of regularized feature visualizations (FV) suggesting what specific neurons have learned. From left to right: FV for the *banana* class in an Inception-based CNN trained on ILSVRC2012, adapted from [272]; FV for the *ball* class in a GoogLeNet (Layer 5a), adapted from [274]; FV for *portrait* (center) and FV got *landscape* (right) classes in an Inception-based CNN finetuned on an art historical dataset to distinguish between portrait and landscape classes, adapted from [271].

inherent flexibility of ACs. In Section II.2.5, we present AC-based image classification baselines, including the experimental setup and the results. In Section II.2.6, we discuss our explainability experiments, including the results from GradCAM++ feature visualizations and our novel SD-AM approach to hypericon creation. In Section II.2.7, we provide a comprehensive discussion of the results, with a focus on contributions, lessons, and future directions. We conclude in Section II.2.8.

II.2.2 Idea: End-to-end Deep Learning Vision

The objective of this section is to establish benchmarks on the novel ARTstract image dataset using state-of-the-art DL models. Specifically, we focus on the utilization of three pivotal architectures known for their excellence in image classification: Visual Geometry Group (VGG) CNN [331], Residual Neural Networks (ResNet) [163], and Visual Transformers (ViT) [109]. Our primary goal is to assess the state-of-the-art DL paradigm for image classification, which involves providing a dataset of images with ground truth labels to a deep neural network. The network autonomously learns relevant features directly from the image data, transforming I_{RAW} into a deep feature representation I_{DL} (see Figure II.2.2). In this chapter, we utilize three readily available models that have been pre-trained on the ImageNet dataset [211]. We fine-tune these networks to extract deep feature vectors as representations of the input images:

$$f_{VGG} : I_{RAW} \rightarrow \mathbb{R}^{512} \tag{II.2.1}$$
$$f_{ResNet} : I_{RAW} \rightarrow \mathbb{R}^{2049} \tag{II.2.2}$$
$$f_{ViT} : I_{RAW} \rightarrow \mathbb{R}^{768} \tag{II.2.3}$$

We separately test performance using each of these feature vectors as input to a classifier head:

$$\hat{y} = \arg\max(p(y_i|I_{\text{DL}}, \theta)) \tag{II.2.4}$$

We evaluate performance by computing accuracy (A), precision (P), recall (R), and F1, (see details in Subsection II.2.5.1).

In our efforts to interpret what the model has learned, we employ two approaches. First, we use the CAM method, which, given $f_y^l(x)$ as the output of the l-th layer of a CNN that classifies the image x with class y, computes a visual explanation map. Secondly, we employ AM, formulated as an optimization problem, such that $\hat{x}$ is an image that maximizes the neuron a_y, responsible for classifying as class y (more details about both approaches are available in Subsection II.2.6.1).

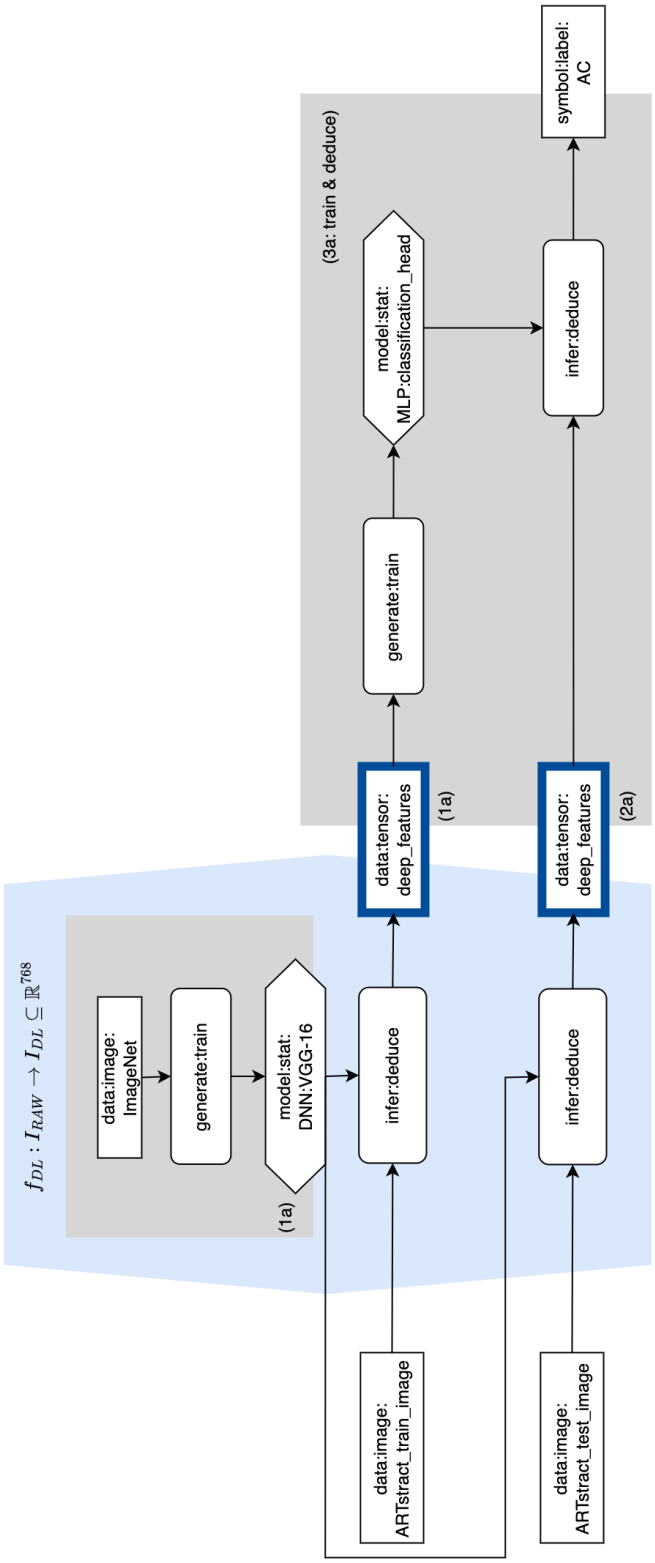


Figure II.2.2: Architecture of the end-to-end deep vision approach to AC image classification.

II.2.3 Related Work

To contextualize our study, we examine the domains of computer vision (CV), explainable AI (XAI), and their intersections with cultural data, to contribute to our understanding of the opportunities and challenges regarding AC image classification on cultural images from ARTstract.

II.2.3.1 Deep Computer Vision for Image Classification

Over the past decade, the field of CV has undergone a profound transformation driven by the advent of DL, specifically powered by CNNs. DL is an important branch developed based on machine learning. It makes full use of the hierarchical characteristics of artificial neural networks to process information and obtain high-level features by learning low-level features and adopting feature combination methods. This approach enables the automatic learning of images and extraction of deep-seated features for tasks such as image classification or regression. A compelling illustration of this paradigm shift from classical CV to the DL paradigm can be seen in image classification, notably catalyzed by the breakthrough of Krizhevsky, et al in 2012 [211]. The introduction of ILSVRC [303], a large-scale image classification challenge on ImageNet [105], marked the introduction of ever-improving image classification models. DL techniques have harnessed the capabilities of extensive data and powerful computing resources to tackle once-considered insurmountable challenges, pushing the boundaries of what is achievable [275]. Since then, DL has consistently outperformed traditional methods in this domain [275].

CNNs represent one of the most widely-used methods in image classification tasks and are the backbone of modern state-of-the-art methods [76]. A CNN is composed of several convolutional layers (see Figure II.2.3a). A convolutional layer learns how to filter an image by learning the filter's kernel. By computing the convolution with the learned kernel over the whole image, the network extracts relevant features for the classification task. CNNs can be classified into three main classes: classical CNNs, inception CNNs, and residual CNNs [76]. Classical CNNs, such as VGG [331], make straightforward use of convolutional layers. Better performances are achieved using deep network models (i.e. networks with a large number of convolutional layers). The use of increasingly deep networks, however, has been shown to increase performances only to a certain extent [76]. To overcome this limitation inception-based methods, such as InceptionNetV3 [348], and residual CNNs, such as ResNet [163], have been proposed.

Recently, following the success of the Transformer architecture [367, 227], transformer-based classification models, such as ViT [398], have been introduced with promising results. Transformers, originally designed for sequential data like

natural language processing, have been adapted to process image patches independently, enabling global context understanding (see Figure II.2.3b). ViT and similar architectures have demonstrated competitive performance, leveraging self-attention mechanisms to capture long-range dependencies in images. Overall, transformers offer advantages such as the ability to model long-range dependencies, adapt to different input sizes, and the potential for parallel processing, making them suitable for image tasks. However, Vision Transformers also face challenges such as computational complexity, model size, scalability to large datasets, interpretability, robustness to adversarial attacks, and generalization performance [246]. For a detailed comparison, please refer to Table II.2.1.

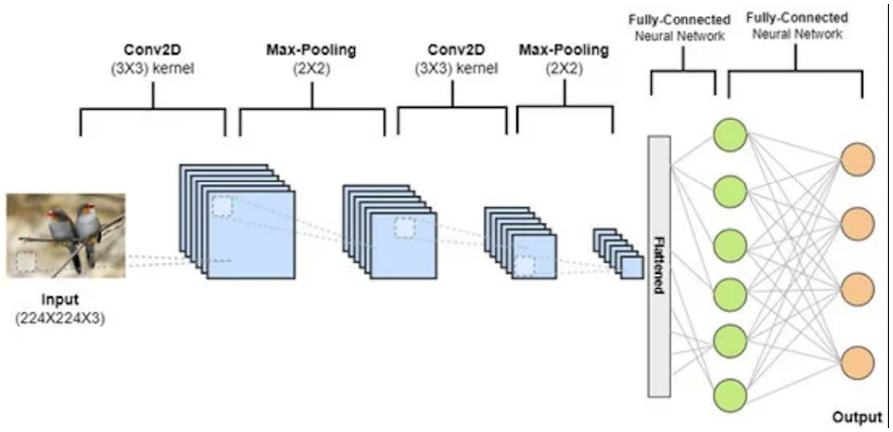
Table II.2.1: Comparison between CNNs and ViT in Image Classification

Feature	**CNNs**	**ViT)**
Architecture	Hierarchical convolutional layers	Self-attention mechanisms over image patches
Flexibility	Limited global context understanding	Global context understanding via self-attention
Performance	Excellent for local feature extraction	Competitive performance
Computational Cost	Lower number of parameters and computational requirements	Higher number of parameters and computational requirements
State-of-the-art	Long-standing dominance in CV tasks	Promising alternative for image classification
Data Hunger	Moderate	Extensive data requirements

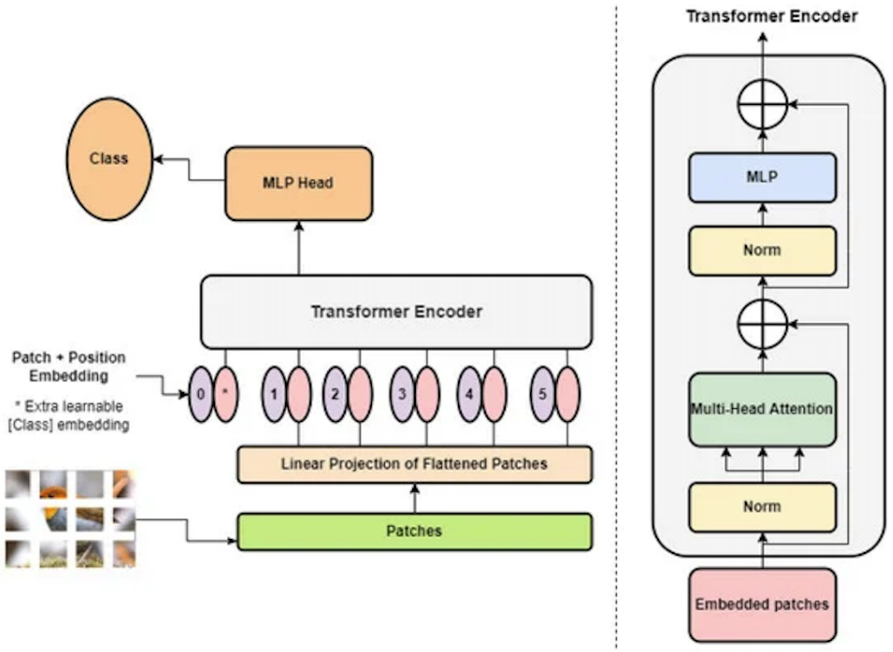
II.2.3.2 Computer Vision and Explainability

In this section, we succinctly describe two of the most commonly used techniques for *post-hoc* explainability of CNN-based CV models [372, 175], CAM and AM.

Class Activation Mapping (CAM). An approach to the visual explanation of image classification models is CAM [406], a method initially proposed to investigate how CNN models trained on classification tasks can generalize on localization tasks. On the XAI landscape, CAM methods are used to highlight the regions of an image that are important in the classification process of a model [175]. Given the output of the last convolutional layer of a CNN that classifies an image, the one that displays the higher spatial resolution when compared to other layers [406], a visual explanation map can be computed by enhancing the response of highly-activated neurons. Different enhancement methods have been proposed: GradCAM [322] and GradCAM++ [74] use the layer's gradient to compute coef-



(a) Example of an architecture of a CNN ViT, based on [310].



(b) Example of the architecture of the ViT, based on [109].

Figure II.2.3: Comparison of CNN and ViT deep neural network architectures.

ficients; XGrad-CAM [126] uses of axioms to avoid the use of heuristic methods; LIFT-CAM [189] proposes an analytical solution to the problem. CAM methods [372, 175] represent one of the main methods used to obtain a visual explanation of image classification models.

Activation Maximisation (AM). A different approach to the visual ex-

planation of image classification models is the AM method [117, 372], where the activation of the classification neurons is exploited to synthesize "prototypical" images of a class. AM is formulated as an optimization problem, where an image is obtained by directly maximizing the activation of one or more classification neurons [265]. The optimization procedure can be expressed in terms of gradient ascent when the gradient is accessible [117, 266, 265]. The resulting image is often hard to interpret from a human point of view. Different regularisation techniques have been proposed to address this issue, such as L_α norm [330] to smooth pixel intensities, Total Variation (TV) [236] to encourage smoothness of the image [265] and gradient enhancing techniques [64]. A different approach to obtain more realistic images and human-understandable images is to synthesize the image using a generator network, as done in DGN-AM[1] [266] by using a Generative Adversarial Network (GAN). This biases the image towards more realistic-looking images that are easier to interpret from a human's point of view. Recently, similar approaches have been used to guide the generation process of Diffusion Models [2] [107]. Such methods have shown promising results in providing *post-hoc* explanations of classification models [182] but have not been applied to the AM process.

II.2.3.3 CV and AC Image Classification

In our systematic survey of works related to AC image classification (see Chapter I.2), we identified a focus on socio-culturally rich images from social media, advertisements, and political contexts, emphasizing the need for nuanced visual data when dealing with ACs [354, 183, 191], and the domain of advertisements stands out based on the prominence of works in this domain [174, 390, 392, 191]. Domain-specific datasets are created in many of the surveyed works, emphasizing the importance of tailored data for AC image classification tasks [174, 354, 183]. Critically, CNN architectures feature prominently in most of the surveyed works, indicating their central role in AC image classification [8, 174, 390, 354]. The surveyed works consistently prioritize F1 scores over accuracy, reflecting the complex nature of AC classification. However, even the highest-performing work, benefiting from a large dataset and textual information, faces difficulties in achieving high F1 scores [354].

II.2.3.4 Computer Vision and Cultural Data

While much of the focus in CV revolves around the analysis of natural, realistic photographs, there is growing research in applying these techniques to art and cul-

[1] https://github.com/Evolving-AI-Lab/synthesizing. Access date: June 2023.
[2] https://github.com/openai/guided-diffusion. Access date: June 2023.

tural historical datasets [341, 298]). One of the central challenges in computational cultural image analysis has been the automated classification of artworks based on categories such as artist, style, or genre. Earlier studies tackled this problem by extracting various handcrafted image features and employing different machine learning algorithms to analyze and classify artworks based on these features [195, 323, 6]. The widespread adoption of CNNs then sparked interest among scholars in their potential to advance humanities and art history research. CV techniques have been applied to analyze extensive painting databases, uncovering patterns and trends [218, 177], to analyze television series [17], and to understand visual trends within digitized Dutch newspapers [382]. Previous work has provided comprehensive surveys of AI's role in art analysis, particularly regarding digitized artwork collections [71].

Cultural images present unique challenges due to their diverse styles and content. Unlike realistic photographs, art, cultural, and historical images often deviate from photorealism. Transfer learning techniques are frequently employed to address these stylistic differences, particularly in handling paintings or historical photographs [410, 382]. Moreover, the objects of interest in cultural images may differ significantly from those in naturalistic photos commonly used in computer vision tasks. Consequently, new datasets and models tailored to the humanities domain are being developed [37]. Research in this area spans a wide range of topics, from detecting unconventional objects like "smelly" objects to identifying zoological species, railway accidents, or musical instruments [409, 342, 333, 305]. Additionally, quantifying subjective aspects of perception, particularly in art images, poses significant challenges. A crucial obstacle lies in creating large-scale datasets annotated with evaluation scores derived from experimental surveys. For instance, Amirshahi et al. [13] introduced the JenAesthetics dataset, which labels artwork images with subjective aesthetic evaluations. Numerous studies have explored computational aesthetics in art, focusing on analyzing statistical properties of paintings [162, 196].

II.2.3.5 Explainability and Cultural Data

Due to the more reflective nature of humanities research, interpretability, explainability, and trustworthiness are core concerns for many scholars that study culture [45]. In a way, computer science has borrowed concepts that contribute to explainability from the humanities domain, such as *provenance*. There is now a vibrant computational provenance community in computer science [106], but this concept originates from (art-)history [260]. Lately, as the humanities domain has a longstanding tradition of source criticism [293], which more recently was expanded to tool criticism [202], there has been a fair amount of attention for biases in datasets and algorithms [381, 272, 335].

II.2.4 Deep Representation Analysis

In this section, we sought to assess the ability of readily available DL (DL) models to encode and represent intricate features relevant to ACs within images. We employed "intraclass similarity" as a metric to gauge the degree to which image representations within the same target class shared common features. The central question addressed in this section was as follows:

> **RQ 1.2.1**: *To what extent do conventional DL representations effectively capture intraclass similarity in images labeled with ACs, as opposed to images associated with concrete concepts?*

Our working hypothesis proposed that the deep representations of images linked to ACs would exhibit lower levels of intraclass similarity when compared to images linked by concrete concepts. This hypothesis is based on the flexibility of ACs and the consequential diversity of images that evoke them, making the capture of intraclass similarity potentially more challenging.

II.2.4.1 Approach

We chose a VGG-16 CNN [331], pre-trained on ImageNet [105], as the reference model due to its widely recognized architecture and established performance in image classification tasks, Despite the availability of models with potentially superior performance, VGG-16's robustness and well-documented features make it a suitable baseline for our study, so we employed it to generate feature vectors for all images in the ARTstract dataset [243]. To quantify the degree of similarity between images within the same AC class, we adopted the cosine similarity metric, allowing us to calculate similarity scores between each image and every other image within its respective class.

To provide context and validate our findings, we identified a benchmark dataset with perceptual content resembling ARTstract but comprising concrete target classes: CIFAR-10. CIFAR-10 is a well-known CV benchmark dataset containing 60,000 32x32 color images distributed across ten distinct classes, with each class containing 6,000 images [210]. Despite the differences in target classes, the visual content of ARTstract images shares similarities with those in CIFAR-10 (see Figure II.2.6). This qualitative comparison between the two datasets allowed us to contextualize our results. To ascertain the significance of the observed differences in intraclass similarity between the ARTstract and CIFAR-10 datasets, we employed a two-sample t-test [88]. The resulting p-values enabled us to determine whether these distinctions in intraclass similarity were statistically significant, indicating meaningful disparities rather than random chance variations. This method

allowed us to establish the robustness of our findings and evaluate the effectiveness of DL representations for AC image classification.

II.2.4.2 Results

Intraclass Similarity ARTstract exhibits notably lower intraclass similarity than CIFAR-10. On average, CIFAR-10 attains 0.644 while ARTstract achieves 0.344. Table II.2.2 compares the intraclass similarity between ARTstract and CIFAR-10, including average similarity values. Figure II.2.7a shows intraclass similarity values for ARTstract and CIFAR-10 classes. The two-sample t-test results show that ARTstract exhibits statistically significantly lower intraclass similarity than CIFAR-10, as evidenced by a basically negligible p-value (rounded to 0.00000) (see Figure II.2.7b), which indicates statistical significance.

Table II.2.2: Intraclass Similarity Comparison between ARTstract and CIFAR-10

(a) ARTstract

Category	Similarity
'comfort'	0.3762
'danger'	0.3304
'death'	0.3664
'fitness'	0.3491
'freedom'	0.3250
'power'	0.3513
'safety'	0.3088
Average	**0.3440**

(b) CIFAR-10

Category	Similarity
'airplane'	0.6346
'automobile'	0.6834
'bird'	0.5973
'cat'	0.6136
'deer'	0.6674
'dog'	0.5819
'frog'	0.6431
'horse'	0.6361
'ship'	0.6767
'truck'	0.7105
Average	**0.6440**

This lower intraclass similarity in ARTstract suggests challenges in discerning shared image characteristics of those tagged with the same AC due to class variance and spatial resolution limitations. It indicates that the level of intraclass similarity in image representations may be influenced by the abstractness or concreteness of the class being considered. Concrete visual concepts tend to have higher intraclass similarity, thanks to the capabilities of CNN-based models in capturing perceptual semantics. Conversely, ACs, lacking concrete visual features, exhibit lower intraclass similarity due to the model's difficulty in capturing semantic nuances. These findings imply that the difficulties in capturing semantic nuances within ACs may lead to classification challenges, particularly in ARTstract.

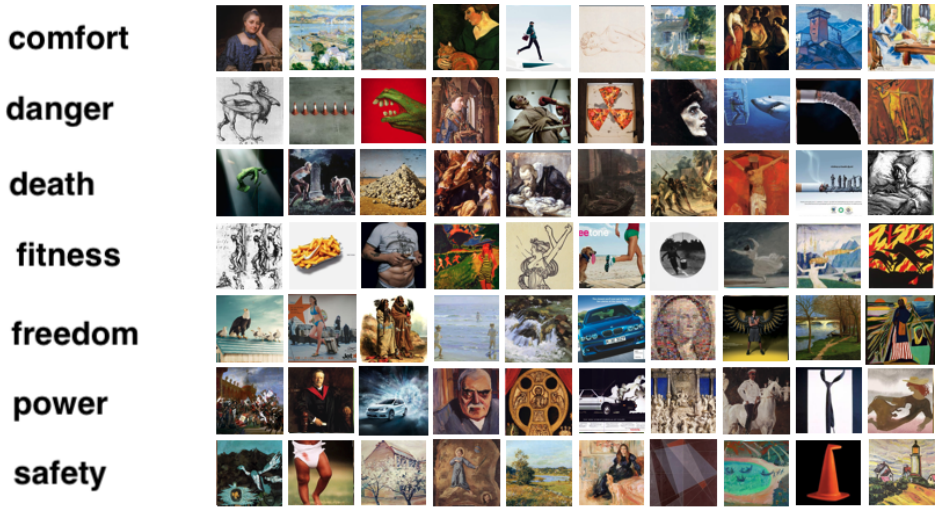


Figure II.2.4: Visual Representation of ARTstract: A selection of ten random images from each of the seven classes in the dataset showcasing the diversity of AC instances.

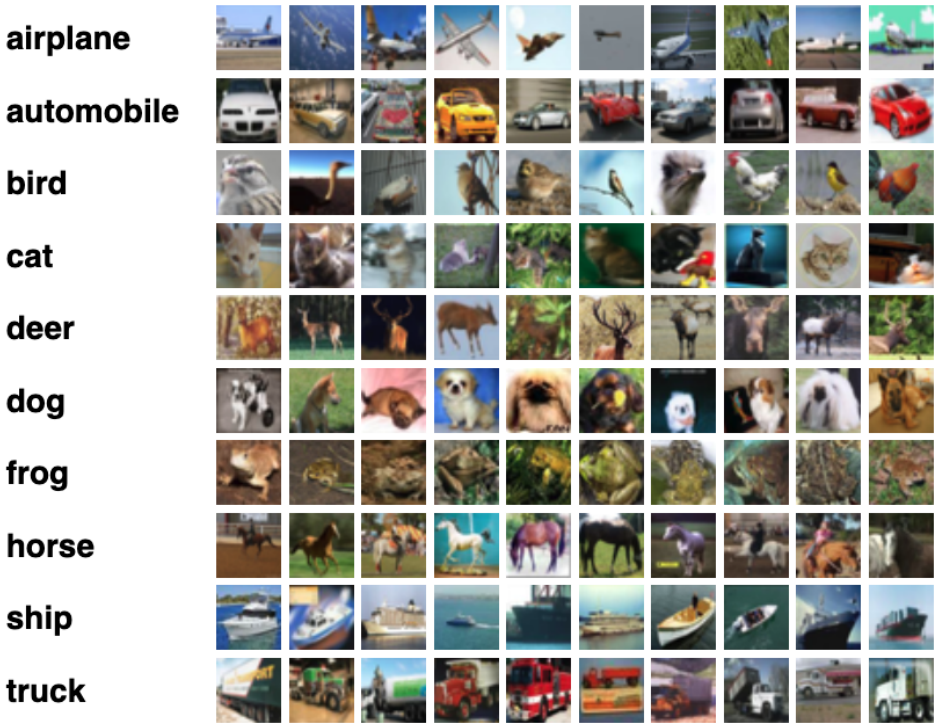
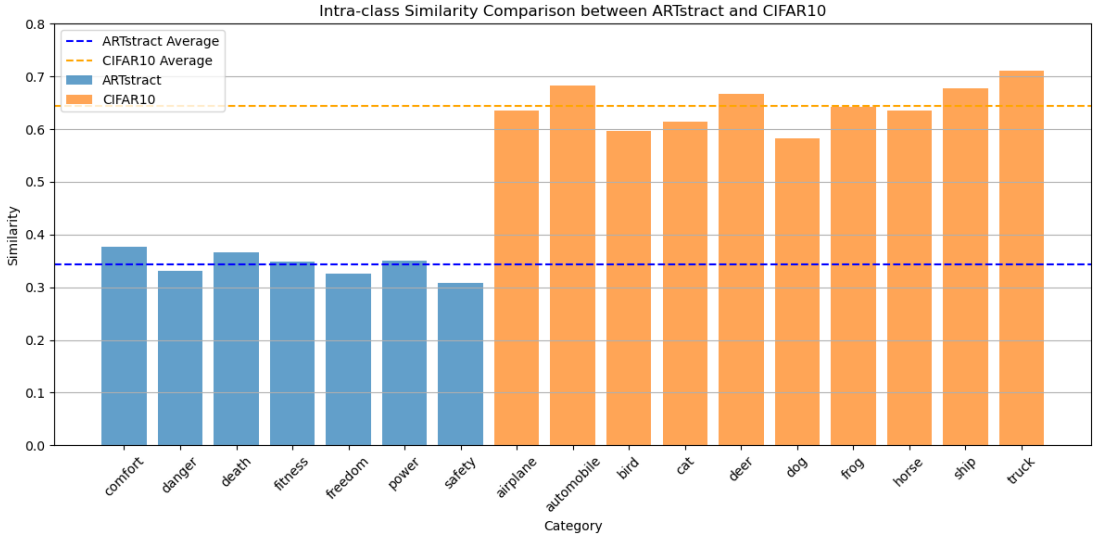
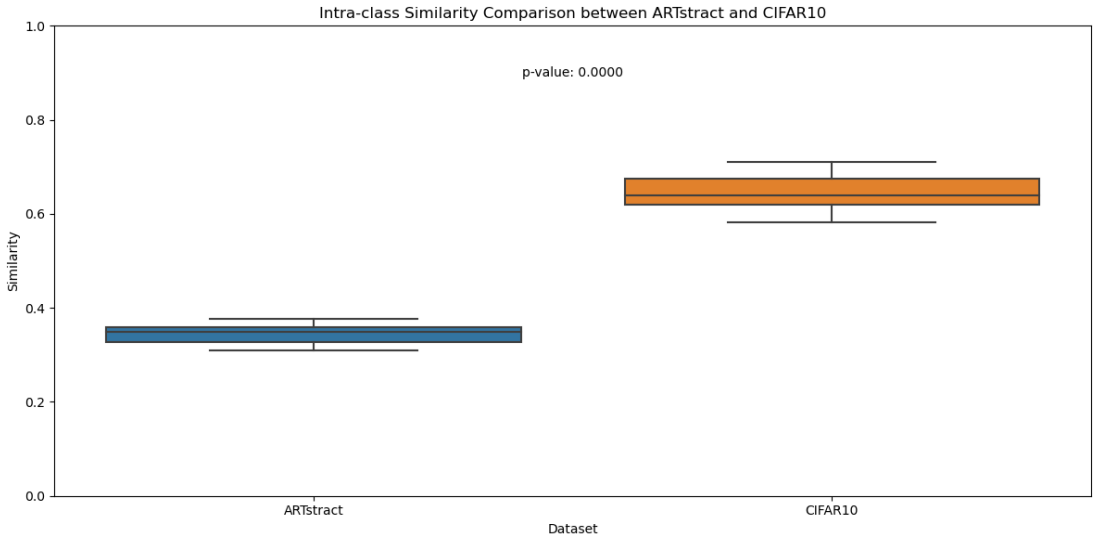


Figure II.2.5: Visual Representation of CIFAR-10: A collection of ten random images from each of the ten classes in the dataset. Source: https://www.cs.toronto.edu/~kriz/cifar.html. Access date: May 2023.

Figure II.2.6: Comparing Intra-Class Similarity: ARTstract vs. CIFAR-10. Even though there is statistically significant higher intraclass similarity for CIFAR-10 classes compared to ARTstract, visually, this substantial difference may not be immediately straightforward.



(a) Intraclass similarity of VGG representations for images belonging to specific classes. The analysis is performed for seven AC classes (in blue) from the ARTstract dataset and ten concrete classes from the CIFAR-10 dataset (in orange).



(b) The extremely low p-value from the t-test indicates that the difference in similarity scores for abstract vs. concrete target classes is statistically significant.

Figure II.2.7: Statistical analysis reveals significant differences in intraclass similarity scores between deep image representations when applied to a dataset of ACs (ARTstract) compared to concrete concepts (CIFAR-10) as target classes.

II.2.5 Deep Performance Evaluation

Building upon the exploration of deep representations in the previous section, the performance of state-of-the-art DL models, including CNNs and Vision Transformer (ViT), for AC image classification was assessed. The goal was to determine how effectively these models could adapt to this specialized task in comparison to their performance in conventional image classification assignments. The central question addressed in this section was as follows:

> **RQ 1.2.2**: *How well do state-of-the-art DL models perform on the task of AC image classification?*

we hypothesized that state-of-the-art DL models would achieve lower accuracy in AC image classification when contrasted with their performance in standard image classification tasks. This expectation arose from the recognition that ACs present unique challenges, and models may need to discern more subtle visual cues than those that might be shared by images tagged with the same concrete tags.

II.2.5.1 Approach

State-of-the-art DL models for image classification were adopted, specifically, two CNNs, VGG-16 [331] and ResNet-50 [163], and a vision transformer (ViT) [109]. These networks had previously been pre-trained on the extensive ImageNet-21k dataset, which endowed them with a strong foundation in feature extraction since pre-training CNNs on large-scale data has been shown to result in more accurate results [203]. The approach was centered on transfer learning and fine-tuning. All three models had the classification head adjusted while keeping the remaining layers immutable and were finetuned for 100 epochs. Additionally, for the two CNNs, we also finetuned the 'whole model', with all layers being unfrozen and fine-tuned.

For evaluation metrics, we follow ILSVRC multi-class formulation [303]. Formally, we estimate the probability $p(y \mid \hat{x}, \Theta)$ where y is among the cluster classes (Y) presented in Section II.1, $\hat{x}$ the input image and Θ is the neural network used to parametrize the probability distribution. Differently than ILSVRC, where the top-k predicted classes are evaluated, we compute accuracy (A), precision (P), recall (R), and F1 defined as

$$A = \frac{1}{\mid Y \mid} \sum_{y \in Y} \frac{TP_y + TN_y}{TP_y + TN_y + FP_y + FN_y} \qquad P = \frac{1}{\mid Y \mid} \sum_{y \in Y} \frac{TP_y}{TP_y + FP_y}$$

$$R = \frac{1}{\mid Y \mid} \sum_{y \in Y} \frac{TP_y}{TP_y + FN_y} \qquad F1 = \frac{1}{\mid Y \mid} \sum_{y \in Y} \frac{P_y \cdot R_y}{P_y + R_y}$$

where TP_y, TN_y are, respectively, the correct prediction of an image in and not in a class and FP_y, FN_y are, respectively, the wrong prediction of an image in and not in a class. We rely on different measures since there is a consistent difference between the number of classes on ImageNet (1000) and the classes on ARTstract (7). Evaluating the top-k results (e.g. top-5 accuracy) would result in over-optimistic results.

II.2.5.2 Experimental Setup

We train each model using an RTX3090 on an Intel i9 CPU with 8 cores and equipped with 128 GB of RAM. We manually adjust and experiment with different hyper-parameters, summarised in Table II.2.3, and train using the Adam optimizer [201], on the whole (unbalanced) dataset split into train, validation, and test sets using an 80:10:10 ratio.

Finetuned Model	Epochs	Batch Size	Learning Rate
ResNet-50	100	32	0.001
VGG-16	100	32	0.001
ViT	100	32	0.001
Whole model ResNet-50	100	32	0.001
Whole model VGG-16	100	32	0.001

Table II.2.3: Hyperparameters used to train the classification baselines.

To further reduce overfitting, we employ standard data augmentation [329] such as random horizontal flips, random color jitter, random rotation, and random crop. For each image in the training set, we resize it to 224x224 pixels, apply a random horizontal flip with a probability of 0.5, a random color jitter with a probability of 0.3, a random rotation with a probability of 0.3, and a random crop of size 20 with a probability of 0.3 and subtract ImageNet mean color. For the images in the validation and testing dataset, we only resize to 224x224 pixels and subtract ImageNet's mean color. Given the described methodology, we can train on the whole ARTstract dataset, minimizing the overfitting due to the skewed distribution of the dataset.

II.2.5.3 Results

In the context of AC image classification using the ARTstract dataset, performances varied from 40% to 51% accuracy (refer to Table II.2.4 and Figure II.2.8 for an overview, and to Section V.1.5.3 in the Appendix, for class-level metrics on VGG-16, ResNet-50, and ViT). The best-performing model, as measured by both accuracy and F1 score, was ViT. Interestingly, for the CNNs, the models that only had the classification head finetuned greatly outperformed the models that had all layers finetuned (see Table II.2.5). This highlights the significance of minimizing extensive fine-tuning of the entire model and the importance of preserving features acquired during pre-training.

Model	Accuracy	Precision	Recall	F1
Head Only (VGG-16)	*0.47*	0.34	0.24	*0.23*
Head Only (ResNet-50)	*0.48*	0.32	0.25	*0.24*
ViT	**0.51**	0.43	0.29	**0.30**

Table II.2.4: Overall performance metrics for the top three best performing models.

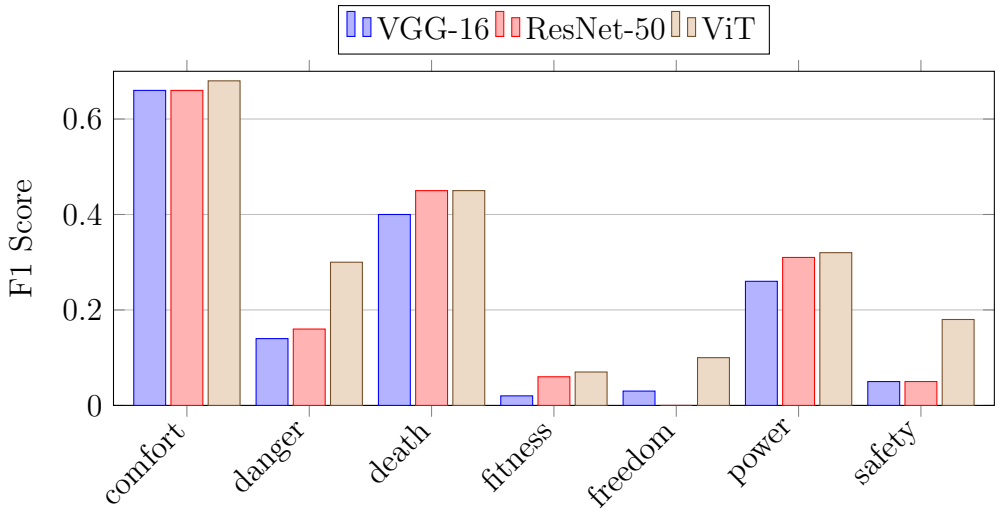


Figure II.2.8: F1 Scores for each of the ACs with ResNet-50, VGG, and ViT.

The comparison of F1 scores between state-of-the-art performance on ADVISE vs. on ARTstract (see Table II.2.6) reveals a significant improvement in the latter. While ADVISE models exhibit F1 scores ranging from 0.13 to 0.15, the ARTstract-trained models consistently achieve higher F1 scores, with values ranging from 0.23 to 0.30. These results suggest that the task definition and use of the ARTstract dataset provides a notable advantage in AC classification over the task defined

Model	Head Only				Whole Model			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
VGG-16	**0.47**	0.26	0.22	**0.23**	*0.40*	0.17	0.20	0.17
ResNet-50	**0.48**	0.33	0.23	**0.24**	*0.43*	0.30	0.21	*0.18*

Table II.2.5: CNN performance on the ARTstract dataset.

Model	F1 Score
VGG-16 (ADVISE)	0.15
ResNet-50 (ADVISE)	0.13
SKG-Sym (ADVISE)	0.14
ASKG-Sym (ADVISE)	0.15
VGG-16 (ARTstract)	**0.23**
ResNet-50 (ARTstract)	**0.24**
ViT (ARTstract)	**0.30**

Table II.2.6: Comparison of F1 scores for state-of-the-art models (trained on ADVISE) and for models trained on ARTstract.

Model	ARTstract	CIFAR-10
ResNet-50	0.48	0.97 [77]
VGG-16	0.47	0.94 [21]
ViT	**0.51**	**0.98** [77]

Table II.2.7: Comparison of model performance between ARTstract and CIFAR-10 Datasets. The CIFAR-10 accuracies, obtained from the cited sources in the table, showcase the models' performance on a well-established benchmark.

and data used in the ADVISE experiments by][191]. The increase in F1 scores indicates that the ARTstract dataset may better capture the complexities and nuances associated with ACs, allowing models to perform more effectively in this challenging task.

A striking point of comparison lies in the performance of the same models on two vastly different datasets—CIFAR-10 and ARTstract (see Table II.2.7). These models, known for their exceptional accuracy in conventional image classification tasks, such as CIFAR-10, exhibit a substantial drop in accuracy when faced with the challenge of AC (AC) image classification on ARTstract. While they achieve remarkable accuracy levels on CIFAR-10, with values ranging from 0.94 to 0.98, their performance on ARTstract is notably lower, with accuracy scores ranging from 0.47 to 0.51. This stark contrast underscores the difficulty of the AC image classification task and the unique challenges it presents.

II.2.6 Deep Explainability Experiments

We wanted to scrutinize the extent to which DL models operate as black boxes when engaged in AC image classification tasks. Additionally, valuable insights that could be gleaned through the application of explainability techniques were unveiled.

> **RQ 1.2.3**: *To what extent do conventional DL models function as black boxes in the context of AC image classification, and what valuable insights can be derived from the application of explainability techniques?*

We hypothesize that traditional explainability techniques would struggle to provide human interpretable explanations for CNN model predictions in AC image classification, but that we might get some lessons from biasing feature visualizations towards human vision by developing "hypericons" to visualize what classification head neurons have learned.

II.2.6.1 Approach

Inspired by [271, 272], we conducted a series of experiments aimed at enhancing the explainability of a single baseline CNN.[3] Our exploration focused on three distinct techniques: CAM, AM, and Stable-Diffused Activation Maximization (SD-AM).

Class Activation Mapping

We are interested in investigating which parts of an image are mostly influencing the classification of our fine-tuned models. We use the CAM method which, given $f_c^l(x)$ the output of the l-th layer of a CNN that classifies the image x with the class c, computes a visual explanation map as

$$CAM(f_c^l(x)) = ReLU(\sum_{i=0}^{N_l} \alpha_k \cdot f_c^l(x)_k)$$

where N_l is the number of channels of the l-th layer, $f_c^l(x)_k)$ is the output of the k-th channel of the l-th layer and α_k is an hyper-parameter of the model. The last convolutional layer is usually taken as l, since it has been shown to display a higher spatial resolution when compared to other layers [406]. GradCAM++ [74]

[3] We chose to focus on CNNs instead of the ViT model due to the distinct nature of ViT explanations and the specific applicability of our chosen methods to CNNs. Within the realm of CNNs, we selected the VGG-16 model for these experiments. The VGG-16 model underwent 1000 epochs of fine-tuning. It is also worth noting that this model was initially trained with an additional class, 'excitement,' derived from a prior version of the ARTstract dataset.

uses the layer's gradient to compute coefficients. Since a precise localization is not the primary focus of our research, but we are instead interested in highlighting the approximate regions of interest for the model for a specific classification, we manually experiment with different methods and decide to rely on GradCAM++ as implemented in `pytorch-grad-cam`[4] [144]. We are interested in investigating which parts of an image are mostly influencing the classification of our fine-tuned models. Since a precise localization is not the primary focus of our research, but we are instead interested in highlighting the approximate regions of interest for the model for a specific classification. We generate saliency maps for images of interest using our VGG-16 models and manually inspect the results (for example, Figure II.2.11 to obtain valuable insights into the classification criteria of the model.

Activation Maximization

To investigate the perceptual topology of the model, we decided to generate AM images for the neuron responsible for a specific class. AM can be formulated as an optimization problem, with the objective function defined as

$$\hat{x} = \underset{x}{\operatorname{argmax}}\, a_c(x) \tag{II.2.5}$$

where $\hat{x}$ is an image that maximises the neuron a_c responsible for classifying as a class c. To generate the AM images, we rely on OmniXAI's[5] [387] feature visualization implementation; we experiment with combinations of different parameters, as described in Table II.2.8.

Parameter	Values
Iterations	300, **400**, 500
Learning rate	**0.1**, 0.01, 0.01
Regularizer	L_1, L_2, **TV**
Regularizer weight	0, -0.05, -0.5, **-2.5**
Fourier preconditioning	**yes**, no
Map uncorrelated colors to normal colors	**yes**, no

Table II.2.8: Activation Maximization parameters supported by OmniXAI [387] library. The best parameters after manual inspections are represented in bold.

[4] `https://github.com/jacobgil/pytorch-grad-cam`. Access date: May 2023.
[5] `https://github.com/salesforce/OmniXAI`. Access date: May 2023.

Stable Diffusion-Activation Maximization

Finally, inspired by the work of DGN-AM [266], where the authors obtain realistic-looking images using a GAN generator, and given the recent success of diffusion models in the automatic image generation task [107, 299, 289], we experiment on the same task by using Stable Diffusion (SD)[6] [299], a diffusion-based image generator model, to synthesize realistic images from the AM (FV) (this method is hereto referred to as SD-AM). Informally, diffusion-based models progressively remove noise from an image using a neural network [166]. The denoising procedure is generally guided by a textual prompt. We exploit this aspect by treating the image produced by the AM method as a noisy image and gradually removing noise from it using SD under two experimental settings (see Figure II.2.9):

- Denoise the AM image without providing any textual prompt (img-to-img);
- Denoise the AM image by also including the class label (e.g. *comfort*, *danger*) as a textual prompt as well (txt-to-img).

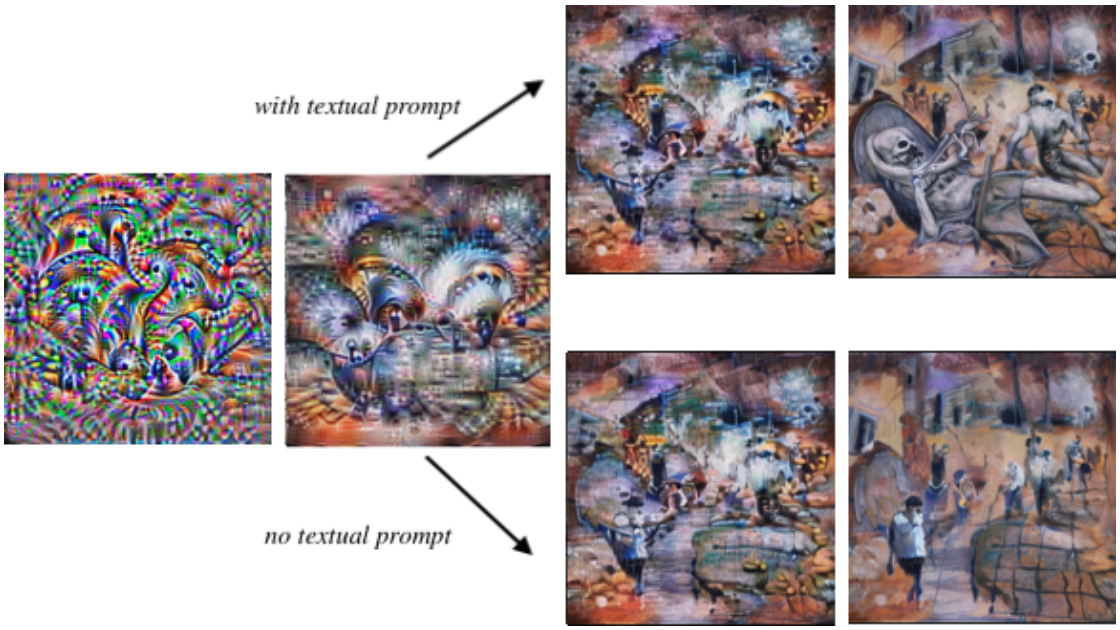


Figure II.2.9: Creation of *death* hypericons via our SD-AM method under the two experimental settings.

[6] https://github.com/CompVis/stable-diffusion. Access date: May 2023.

II.2.6.2 Results

Class Activation Mapping

(a) Original

(b) Comfort

(c) Freedom

(d) Safety

Figure II.2.10: GradCAM++ for three different classes computed using the fine-tuned VGG-16 model on *Triumph of the Virtues over the Vices* painting by *Paolo Fiammingo*, circa 1592. Oil on canvas, dimensions 16.5 × 221 cm (6.4 × 87 in). Image sourced from Wikimedia Commons, originally from Sotheby's auction in London on 6 July 2011.

(a) Original Work.

(b) Heatmap for *freedom* in Original Work. (c) Heatmap for *freedom* in Derivative Work 1. (d) Heatmap for *freedom* in Derivative Work 2.

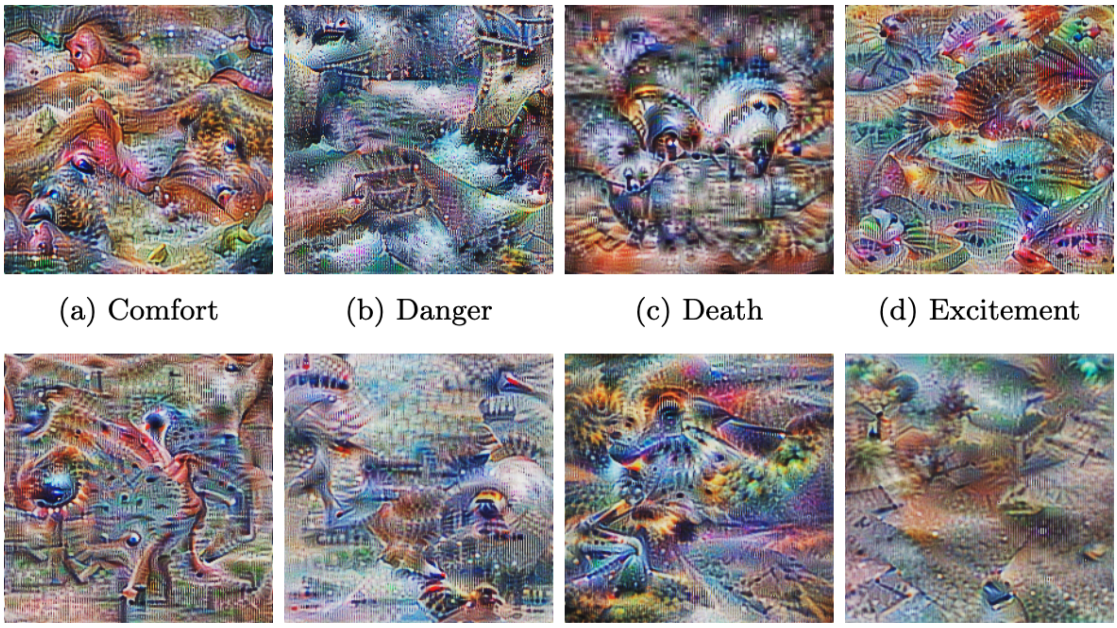
Figure II.2.11: Top: *Liberty Leading the People*, oil painting by Eugène Delacroix, 1830, dimensions 260 cm × 325 cm, Louvre, Paris; image sourced from Wikimedia Commons. Bottom: GradCAM++ computed for *freedom* on (1) the original painting, and two (2-3) derivative paintings. The three heatmaps show similar activation areas for the class of *freedom*.

We rely on GradCAM++ to experiment in identifying which parts of certain images are valuable from the point of view of the fine-tuned classification models.

The saliency maps for images of interest previously unseen by the model are done by activating specific classes in our VGG-16 model. We present some results on images previously unseen by the model in Fig. II.2.10 and Fig. II.2.11). The latter figure presents the resulting heatmaps for *freedom* on Delacroix's iconic painting as well as on two derivative works inspired by it, which re-interpret Delacroix's within the context of Hong Kong protests.[7]

Activation Maximization

We applied the logic of [271] to create feature visualizations for optimizing the activation for each of the classifier neurons for each of the 8[8] target classes for the VGG-16 based model (Fig. II.2.12).



(a) Comfort (b) Danger (c) Death (d) Excitement

(e) Fitness (f) Freedom (g) Power (h) Safety

Figure II.2.12: Activation Maximizations (AM), also known as Feature Visualizations (FV) for each of the 8 target classes using the finetuned VGG-16 method.

[7] "Liberty Leading the People of Hong Kong" by Frederic Bussiere, group exhibition "The Art of Resistance", Kong Art Space, Hong Kong, 2019, Digital collage, printed on canvas. 90 x 90 cm https://www.behance.net/gallery/90838377/Liberty-Leading-the-People-of-Hong-Kong-collage and "Our Vantage" by Harcourt Romanticist https://www.instagram.com/p/B2EQ1FPnDh1/. Access dates: May 2023.

[8] A reminder that the VGG-16 model used for these experiments was trained on an earlier version of ARTstract with an additional class, *excitement*.

(a) Hypericons for *comfort*, *danger* and *death*, resulting from our SD-AM method. The figure shows two hypericons per concept: on the left, the prompt-agnostic one (denoised without any explicit prompting for the target class), and on the right, prompt-guided (denoised with the target class label as input).

(a) Hypericons for *fitness*, *freedom*, *power* and *safety* resulting from our SD-AM method. The figure shows two hypericons per concept: on the left, the prompt-agnostic one (denoised without any explicit prompting for the target class), and on the right, prompt-guided (denoised with the target class label as input).

Stable Diffusion-Activation Maximization

With the stable diffusion-based image generator model, we synthesized realistic images from the ones obtained using the AM method. Examples with denoising detail from both experimental settings (with textual prompt and without textual prompt) are hereby presented: for *death* and *freedom* (see Fig. II.2.9). The final hypericons obtained for our 7 classes of interest are presented in Figure II.2.13a

II.2.7 Discussion

In this study, we critically examined the prevailing trend of automating high-level visual reasoning through DL, using the ARTstract dataset as our case study, where cultural images are associated with ACs. Our investigation encompassed a multi-faceted approach, involving an analysis of deep representations through intraclass similarity metrics, the training and evaluation of state-of-the-art DL image classification models, introduced as baselines, and the implementation of explainability techniques to try to "open" the black box models. Our particular emphasis was on unraveling the knowledge sought and encapsulated by CNNs. This section offers a discussion of the presented results and an in-depth exploration of the utilization of explainability techniques, including the introduction of our novel approach, SD-AM, which facilitates hypericon creation. Through these comprehensive efforts, we aimed to illuminate the challenges and potentials inherent in socio-cultural visual reasoning and emphasize the pivotal role of explainability in mitigating biases and ensuring the fairness of AI systems.

II.2.7.1 Deep Representation

The results indicate that there are differences in intraclass similarity between ARTstract and CIFAR-10 classes, which are highly unlikely to occur by random chance alone. The results shed light on the dependence of intraclass similarity in image representations on the concreteness or abstractness of the target class. The CNN-based model effectively captures shared features among images within concrete label target classes, resulting in higher intraclass similarity. Conversely, images associated with the same ACs exhibit lower similarity, signifying fewer shared features. The remarkably low p-value underscores the significance of these findings. This suggests that the model's spatial resolution struggles to capture semantic nuances, leading the VGG-16 ConvNet to perceive all ARTstract images as similar, regardless of their specific AC. These results underscore the inherent challenges of representing ACs within traditional DL models. The lower average intraclass similarity in ARTstract highlights the difficulties posed by class variance and spatial resolution limitations when compared to concrete classes. Further exploration and

analysis of these distinctions may yield valuable insights into DL models' ability to capture both abstract and concrete semantic information.

II.2.7.2 Deep Performance

The performance of models in ACimage classification on the ARTstract dataset displayed a notable range, with accuracy values spanning from 40% to 51%. The Vision Transformer (ViT) emerged as the top-performing model, achieving the highest metrics in both accuracy and F1 score. This finding underscores the potential of transformer-based architectures like ViT in the field of image classification. The underlying technology of Vision Transformer (ViT) is discussed in [398], and it has been shown to detect features comparable to CNNs [287]. However, it is worth mentioning that training or fine-tuning such models, including ViT, is problematic without a proper large-scale dataset [339].

When considering the performance of CNNs we found that models that underwent fine-tuning on only the classification head significantly outperformed those with all layers finetuned (see Table II.2.5). This result highlights the importance of preserving pre-trained features and minimizing extensive fine-tuning, a practice that can potentially enhance model performance in this specific task. This also suggests that the use of pre-trained models, such as those from ImageNet, offers a promising foundation by leveraging their learned visual features. Training a model for AC classification allows us to investigate to which extent low-level perceptual features can be used on this task. This allows us to better understand which insights, if any, from these methods can be used to more effectively deal with the task.

The comparison of F1 scores between models trained on the ARTstract dataset and those from related works, such as ADVISE [191], unveils a significant performance improvement in the former. Specifically, while the ADVISE models achieved F1 scores ranging from 0.13 to 0.15, our ARTstract-trained models consistently exhibited higher F1 scores, falling within the range of 0.23 to 0.30 (see Table II.2.6). This outcome underscores the substantial advantages afforded by the task definition and use of the ARTstract dataset in the context of ACimage classification, surpassing the task definition and data employed in the ADVISE experiments. These findings illuminate the dataset's effectiveness in capturing the intricacies of socio-cultural visual reasoning, positioning it as a valuable resource for advancing the state of the art in this domain.

A significant point of comparison arises when contrasting the performance of these models on the ARTstract dataset with their performance on the CIFAR-10 dataset (see Table II.2.7). While the models demonstrated exceptional accuracy in traditional image classification on CIFAR-10, achieving values between 94% and 98%, their accuracy scores dropped significantly when confronted with the chal-

lenge of AC image classification on ARTstract, ranging from 47% to 51%. This striking disparity underscores the unique difficulty posed by AC image classification and the distinctive challenges it presents compared to more conventional image classification tasks. They also highlight the importance of dataset choice and task definition in shaping model performance. Even though more complex training procedures can be employed [192] (e.g. training only a subset of the total number of convolutional layers), we argue that the intraclass variance displayed by ARTstract hardly allows any significant improvement from a quantitative point of view [42, 328].

Figure II.2.10 illustrates some of the challenges that deep models may encounter in AC image classification. GRAD-CAM++ results show that the concept of *freedom* is linked to a region containing a weapon, while *comfort* and *safety* are associated with different areas. When an image contains regions associated with different ACs, the model may struggle to determine which regions and ACs are the most significant. In this example, the system may find it challenging to decide whether the presence of weapons in the image should be sufficient for classifying it as *freedom*, or if, conversely, the identification of regions relevant to *comfort* in the top-right of the image may overshadow it.

Overall, the performance results of this study bring to the fore the wicked nature of the problem of automatically detecting ACs within computer vision. The proposed AC image classification baselines show relatively low performances when compared to other CV tasks on art images, such as style, genre, or artist classification [350, 70]. The results shown in Table II.2.4, however, are similar to the results obtained by models that use a similar amount of images on a radically different set of labels compared to ImageNet [264]. The complexity of detecting ACs might hence stem from the relatively open definition of each AC, which does not explicitly account for their polysemy and association to vastly varied visual data (as seen in the example of *danger* in Figure II.1.2). The shallow representation obtained using a CNN-based method is not able to generalize enough to capture the ambiguities of such definitions.

II.2.7.3 Deep Explainability

The fact that even with low performances, high-level, subjective tasks like AC image classification are increasingly automated underscores the pressing need for interpretability. The inherent complexity, contextual dependencies, and subjective nature of ACs demand a level of transparency and explainability not only for improving model performance but also for gaining insights into the underlying reasons for model decisions. By uncovering the visual cues and features that contribute to classification outcomes, interpretability facilitates informed refinement and adaptation of algorithms.

Insights from Traditional Explainability: CAM

We experimented with a CAM-based method as it is the most well-known explainability technique to uncover the decision-making mechanisms of CNN-based models. These insights were pivotal in shedding light on the complex processes underlying our models' decisions. To explore the reasons behind our models' classification of unseen images, we employed GradCAM++ to pinpoint valuable image regions. Using the finetuned VGG-16 model, we activated specific classes for previously unencountered images. This approach provided valuable perspectives into the model's decision boundaries. Figure II.2.11 displays the results for the class of *freedom* in Eugène Delacroix's painting "Liberty Leading the People" and two derivative works inspired by it. In this example, both the original work and the two derivative works localize the *freedom* class in the same area (i.e. where the flags and raised hands are located). We can hence discern how the model identifies specific visual cues that evoke the AC. For instance, the emphasis on flags within the tested images suggests a connection between the concept of *freedom* and symbols of nationhood or political expression. This shows that, indeed, the model can identify perceptual components that are relevant to the image. Such perceptual components, however, are to be considered as lower or mid-level image features in the context of AC classification problem. This suggests that these types of perceptual semantics (such as objects) can be intermediaries that help bridge the gap between raw pixels and ACs.

Another aspect that emphasizes the usefulness of such a technique is its consistent application on the same image, but to localize important regions for different classes. Figure II.2.10 presents the results on three different classes in the painting *"Triumph of the Virtues over the Vices"*. These experiments identify parts of the image in which the model focuses for the image's classification as a selected class. For instance, *comfort* is localized near the figure sitting in a relaxed position on a comfortable couch. In contrast, *freedom* is concentrated around the area of the painting with angels, clouds, and a raised sword. These findings suggest biases in the ARTstract dataset, potentially stemming from *freedom*-tagged images being biased towards images depicting elements like raised swords or flying agents such as birds or angels.

Overall, these results showcase the effectiveness of CAM-based methods in identifying valuable regions in images for classification models, thereby highlighting potential biases in the dataset and providing insights into how the model perceives and processes images. However, the results also underscore that while CNNs are aware of statistical correlations, these correlations may not always align with human perspectives. Despite providing valuable insights into classification processes and the identification of ACs in images, the shallow representation achieved by the model can yield false evidence. Furthermore, the lack of robustness, particularly

against adversarial attacks, poses a significant concern for the interpretability of classifiers [9]. In conclusion, these findings stress the need for further research to enhance the accuracy and robustness of classification models when addressing ACs in the realm of cultural images.

AM: Distributed Reality, Perceptual Bias, and Feature Visualization

We see the regularized feature visualizations for each of our eight AC targets (shown in II.2.12) as examples of how distributed reality (in terms of manifestations and perspectives) can get collapsed into one 2D image. When they are learned and represented by a CNN, concepts are "dissolved", or "entangled", losing their spatial coherence, and thus "it is no surprise that feature visualization images will reflect different manifestations of, and perspectives on, an object, akin to Cubist paintings" [273, pg. 1301]. We see ACs as a prime example of how visual concepts get dissolved in ways that are practically unintelligible to humans. Additionally, the fact that the images in the referenced figure are regularized means that we already introduced a syntactic bias to guide the manifestations into a textural landscape closer to what we visually comprehend. With this syntactic optimization, most of the FVs in Figure II.2.12 are still relatively humanly incomprehensible (no noticeable objects or otherwise *legible* items are very visible). An exception may be the case of the FV of *fitness*, in which certain edges seem to resemble a human figure playing with some sort of ball (see Figure II.2.15).

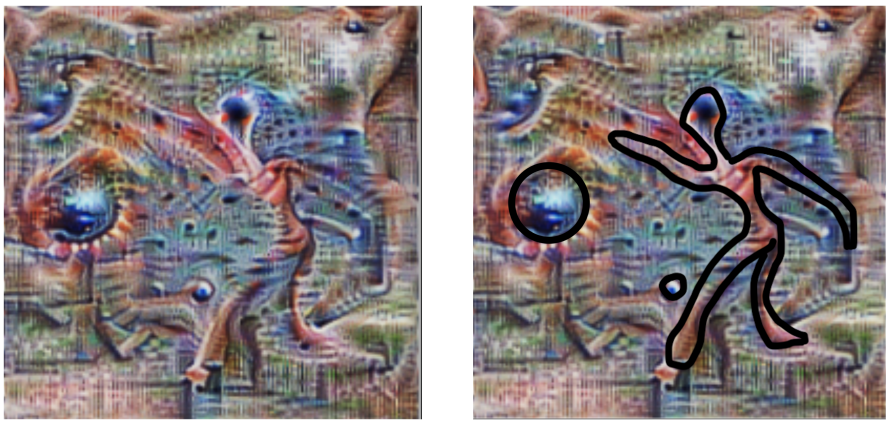


Figure II.2.15: The feature visualization (FV) of the *fitness* class using VGG-16 finetuned for 1000 epochs (left) resembles a human figure playing with a sort of ball (right).

SD-AM: Denoising and the Big Trade-Off

Because moving from the entangled state to a state of higher semantic interpretability for humans requires introducing more constraints, in this work we decided to combine the feature visualizations produced with the AM method with Stable Diffusion (SD-AM) (see Figure II.2.16). Despite some work on the generation of counter-factual explanations [182, 397], the investigation of diffusion models trained on large-scale data, such as Stable Diffusion, as tools that allow a better understanding of classification models is a novel approach. While past attempts, such as combining AM with GAN models [266] have shown promising results, we argue that the intrinsic dependency of GANs on a classification model (the discriminator) can potentially bias the generation process toward results that are harder to automatically classify for syntactic reasons (i.e. the color distribution) rather than for semantic reasons. This is especially true for ACs, where the difference between classes, for instance *danger* and *power*, depends on the semantic content of the image alongside the interpretation of the AC that the network inductively learns. The use of diffusion models overcomes this limitation by design. An image is not generated to fool a discriminator, but rather to remove the noise that makes an input image hard to interpret for humans. While the experiments on FV (Figure II.2.12) provide little insight into the perception of an AC by the model, the denoised version of such images, especially those that are textual prompt guided, (Figure II.2.13a and II.2.14a) let prototypical versions of an AC emerge.

However, it is important to note that the images produced are confined to the latent space of the specific diffusion model employed, similar to the one argued regarding GANs [273]. These images do not reflect the perceptual topology of the analyzed CNN, but they rather replace the elements that are hard to interpret with what the model perceives as human-like, essentially filling the gap between AM images and interpretable hypericons. As such, SD-AM is to be taken as an explorative approach towards easier interpretation of AM images. Further research is required to investigate whether this approach can be directly incorporated into the extraction of saliency maps, to obtain human interpretable results while minimizing the influence of SD.

Perhaps the most critical contribution of the studies that inspired our work [271, 272] is their discussion of the problem of perceptual bias in machine vision systems, which can only be overcome by shifting toward different biases. As they discuss, any constraint added to the optimization process for feature visualizations moves the images further away from showing the actual perceptual topology of a CNN, unveiling the trade-off between representational capacity and legibility of feature visualization images. Their work highlights that feature visualization is one way to achieve forced legibility but also presents a dilemma that the representational capacity of feature visualization images is inversely proportional to

their legibility. Feature visualizations that show "something" are further removed from the actual perceptual topology of the machine vision system than feature visualizations that show "nothing."

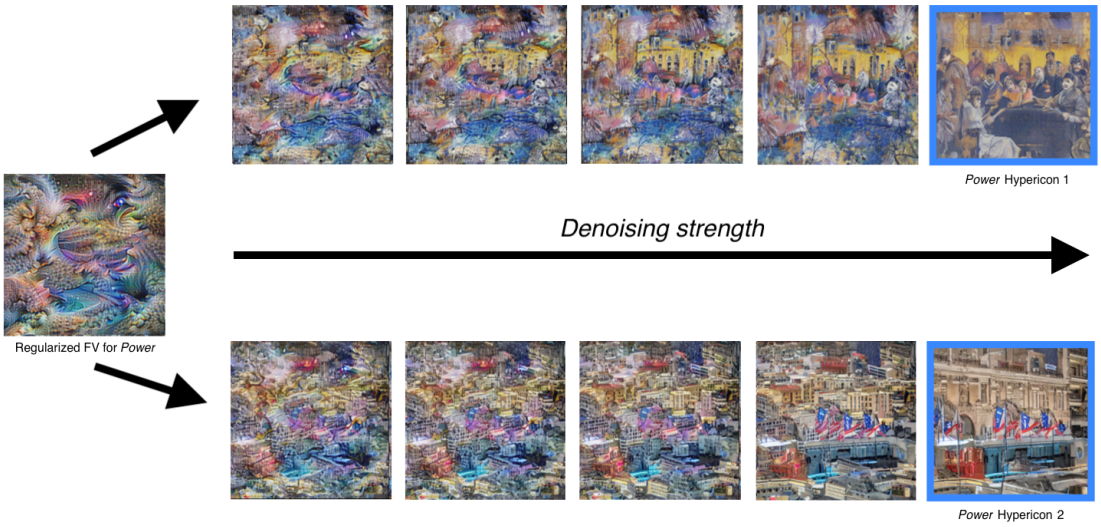


Figure II.2.16: Starting with the regularized FV for the *power* class, the progression from left to right illustrates a gradual escalation in denoising strength, resulting in images with enhanced human interpretability. This procedure was performed twice, yielding distinct SD-AM "hypericons" for the *power* category (blue borders), both originating from the same initial regularized FV. It is noteworthy that the denoising process was executed **independently of any textual prompt**, thereby ensuring that the process remained entirely oblivious to the correlation between the FV earmarked for denoising and its status as representing "power".

While keeping in mind this trade-off, they also note that we can treat resulting hypericons as valuable tools for visual interpretability. Hypericons, such as the ones presented in Figure II.2.13a and discussed in the previous section can be used in combination with the original datasets (in this case, ARTstract) to enable the identification of interesting patterns, especially when treating them as [253, p. 49] suggests:

> The metapicture is a piece of moveable cultural apparatus, one which may serve a marginal role as illustrative device or central role as a kind of summary image, what I have called a 'hypericon' that encapsulates an entire episteme, a theory of knowledge.

As such, in addition to aiding the understanding of the perceptual topologies of CV models, feature visualization images can be studied as concrete representations of cultural knowledge defined by the lenses and tags fed into the CV systems. We hope that our analyses in this case study can function as evidence that exactly this "subjective" nature of feature visualization images is what can make "visual explainability useful in computer vision for art" [271].

A crucial point is that the SD-AM can lead to relevant and semantically meaningful hypericons even without any textual prompt, i.e., without being guided or biased by the class corresponding to the input AM image. For example, in Figure II.2.16, we present results of applying promptless SD-AM to obtain hypericons related to the *power* class. We denoised the extracted and regularized AM by gradually increasing the intensity (weight). By increasing the intensity of the denoising process, we can control the number of transformations applied by SD to obtain an image that is perceived (by the model) as closer to its original training data. Critically, as seen in Figure II.2.17, this process effectively converges towards more human-intelligible hypericons, which resemble real instances of artworks from the corresponding class present in the original ARTstract dataset both visually and semantically.

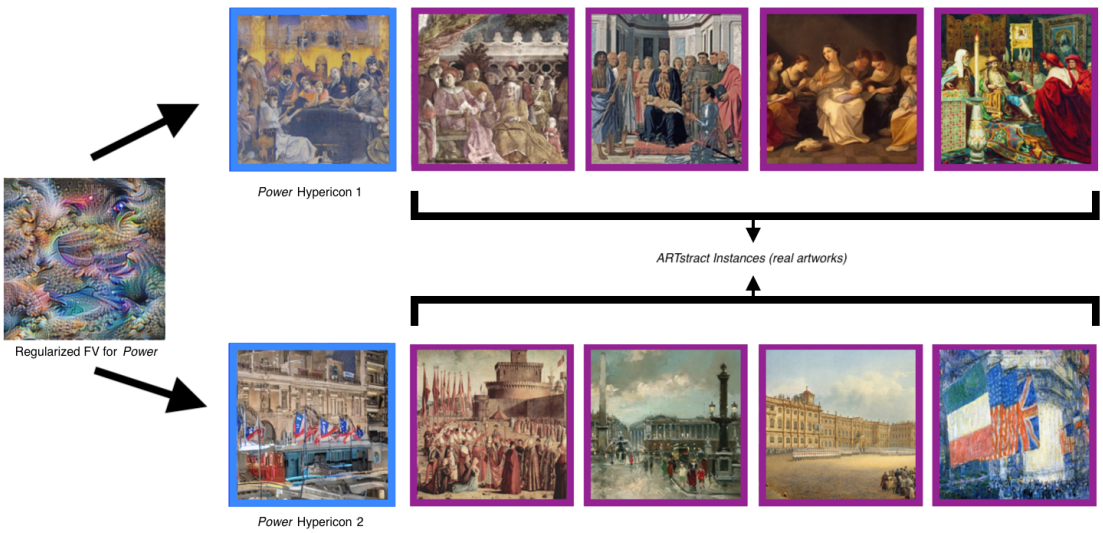


Figure II.2.17: Comparison of the synthetic SD-AM "hypericons" for the *power* class (with blue borders) with manually selected with real instances from ARTstract (with purple borders). These real images from ARTstract are tagged with *power*, they were selected because of their visual and semantic similarity to the hypericons.

Hyperpop Hypericons

The proliferation of images in modern mass media has reached unprecedented levels, with social media platforms alone hosting billions of images every day, resulting in a visually abundant contemporary culture where individuals are bombarded with heterogeneous visual data. This phenomenon characterizes the post-modern era, where users are overwhelmed by an abundance of information that is not curated [181], making it increasingly important to become "lookers" in addition to being readers [334]. The rise of hyperpop and meme culture, favored by the newest generation of technology users searching for sense amidst the chaos, is symptomatic of the current historical moment. As Vassar [366] suggests, hyperpop serves as an attempt, "suited to the psyche of the six-hours-of-screen-time-a-day individual", to strive to find some semblance of meaning amidst the disarray, by compiling a vast field of disparate meanings until they reach some semblance of accord.

Intriguingly, an uncanny resemblance between the SD-AM hypericons and hyperpop artworks surfaces, as illustrated in Figure II.2.18. Within this visual dialogue, a profound parallel emerges as both categories exude dissolved yet intricately collapsed visuals. These striking visuals offer a blend of object fragments and hues, all while boundaries remain indistinct. This parallel beckons us to explore the intersections of aesthetics and cognition. Intriguing questions arise—does the act of meaning collapse within the hypericons mirror the cognitive underpinnings of the hyperpop aesthetic? Could it signify a convergence of overstimulated sensibilities, reminiscent of the torrent of data processed by our models? Analogously, just as our hypericons weave significance from intricate data, the hyperpop aesthetic might mirror the cognitive fabric of contemporary generations, offering a fresh lens for interpreting the avalanche of visual content.

Explainability Lessons for DH

We can assume that detecting and correcting bias of CV systems in the context of cultural data and Cultural Heritage (CH) will mostly happen in a *post-hoc* manner, i.e., after a system has been deployed in real-world situations. This is because many models based on similar patterns have already been used in real-world applications, especially in the digital humanities. We believe that the integration of interpretability into CV-based systems in the cultural heritage (CH) field has not received enough attention. This study stands as proof that digital humanities (DH) initiatives can act as valuable arenas for both probing the limits of established CV explanation techniques and pioneering novel methodologies. DH projects, with their interdisciplinary focus and emphasis on interpretation, offer a unique opportunity to combine technical methods with hermeneutic work to develop systems that are interpretable-by-design. We envision our work as one

(a) Eight human-made artworks representing the recent "hyperpop" aesthetic, pulsating with collapsed visuals and meanings, symptomatic of an era of information abundance and visual saturation. Top row: artworks by Claire Barrow (2020-2023) [178]. Bottom row: artworks by Mikey Joyce (2020-2023) [179].

(b) Eight SD-AM hypericons crafted for each of the 8 AC classes, emblematic of our work, revealing a surprising resemblance to the aesthetic of hyperpop artworks.

Figure II.2.18: Visual convergence of (II.2.18a) hyperpop aesthetics and (II.2.18b) SD-AM hypericons, a juxtaposition that invites contemplation on the parallels between the rapid pace of modern media consumption and the massive data flow into DL models.

of many DH projects that can contribute to the broader development of more transparent and understandable computer vision systems. With their diagnostic capability, the tools developed in XAI are exciting both to the technical disciplines for improving the systems they develop, and to fields such as Digital Humanities with alternative paths for thinking about the kind of work they do (e.g. by interrogating through explainable methods the way that a system has classified certain cultural objects) [44].

II.2.7.4 Future Directions

To address the existing limitations and expand the scope of the study, several avenues for future research are identified:

- **Further Pre-training and Fine-tuning**: Further pre-training or finetuning models, initially trained on ImageNet, on other available art datasets to achieve specialization in AC detection could be useful. Further fine-tuning of the final convolutional layers on ARTstract can enhance classification performance [192].

- **Utilization of Knowledge Graphs (KGs)**: Hybrid methods using KGs as background knowledge offer an opportunity to enhance classification accuracy and robustness [238, 400]. Additionally, exploring KGs for better explainability by incorporating semantic information into CV systems can improve transparency and interpretability.

- **Prominent Region Detection**: Further experiments in detecting specific regions where an AC can be prominently identified can enhance classification accuracy. Systematically evaluating regions identified using CAM-based methods and state-of-the-art techniques in semantic segmentation can provide deeper insights into image interpretation [265, 255]. Running unsupervised attention mechanisms to locate crucial image areas can aid in bounding box designation [175].

- **Improving AC Classification Explainability**: Conducting diverse analyses such as color palette analysis, co-occurring object detection, and exploration of visual cues like texture and style can enhance classification explainability. Additionally, exploring pose detection and other visual cues can further improve interpretability.

- **Human Performance Benchmark**: Assessing AC image classification methods against human performance is crucial for gauging their real-world utility. Designing task-based user studies to mimic AI model tasks can offer

valuable insights. Diverse participant backgrounds should be included for a comprehensive assessment.

- **Explainability User Assessment**: Evaluating the explainability of AI systems is vital for establishing trust and transparency. Future research should conduct user studies to assess the clarity, comprehensibility, and usefulness of various explainability methods, including feature visualizations, hypericons, and diffusion processes.

- **Refining Task Definition and Metrics**: Single-label multi-class classification may not fully capture the complexity of associating multiple ACs with a single image. Proposed refinements include exploring ranking-based tasks, prioritizing reasonability over objectivity in evaluations, and developing evaluation metrics considering semantic relationships between AC classes.

II.2.8 Conclusions

In our pursuit to address the intricate challenge of automatically classifying images based on evoked ACs, we have experimented with DL models on the novel ARTstract dataset, and used these as a lens through which we delve into the realm of explainability. This work unravels the role of end-to-end vision models in complex high-level visual tasks, establishing benchmark model performances for AC image classification within the ARTstract context. Additionally, we combine traditional and novel explainability techniques to better understand model behavior and predictions. With SD-AM, by harmonizing AM with diffusion models, we create synthetic "hypericons" that compellingly visualize the profound transformation of AC meanings as captured by deep networks into singular images. Our study resonates with the burgeoning demand for interpretability in CV systems, especially within the CH domain and the realm of socio-cultural-cognitive visual understanding. We accentuate the significance of recognizing biases and forging connections between the technical and humanistic dimensions, advocating for unconventional pathways to extend hermeneutics. In conclusion, this chapter calls for the explicit integration of explainability into the fabric of CV-based systems that attempt to address high-level visual challenges. This integration is vital to ensure the dependability and credibility of these systems in the evolving landscape of art, culture, and technology. It beckons us to challenge the binary boundaries that prevail in CV, advocating for a more culturally situated, humanistic, and multifaceted perspective.

Part III

Minding the Gap with Cognitive Intermediaries

Chapter III.1

Automating Abstract Concepts' Acquired Embodiment

Summary The results obtained in Chapter II.2 provided valuable insights into the potential link between specific perceptual elements, such as depicted objects like swords or flags, and the visual evocation of ACs like *freedom*. These findings align closely with the concept from cognitive science referred to as *acquired embodiment* [167] (as discussed in Chapter I.3, Section I.3.5). Acquired embodiment serves as a mechanism for inferring sensory-perceptual attributes associated with ACs, particularly those that may lack strong sensorimotor connections. ACs acquire sensory-motor features based on their shared linguistic contexts with concrete concepts. These findings collectively suggest the presence of visual data descriptors, potentially anchored in specific depicted objects, which have the potential to bridge the gap between raw pixel data and ACs. Building on this foundation, our research in this chapter endeavors to translate the cognitive concept of acquired embodiment into a computational method capable of identifying concrete sensory features that visually ground ACs within images in a data-driven manner. Our method focuses on representing ACs as multimodal frames by integrating sensory perceptual and semantic data extracted from images and their metadata. This exploration hinges on identifying patterns of co-occurrence involving dominant colors, concrete objects, and depicted actions in images tagged with specific ACs. We define a conceptual model and introduce a novel ontology for the formal representation of ACs as multimodal frames that can integrate the extracted information. To illustrate the viability of our approach, we conduct experiments using the Tate Gallery's collection as an empirical basis, demonstrating the proof of concept. Furthermore, we discuss potential avenues for future research and provide access to all associated software, data sources, and results.

III.1.1 Introduction and Background

In this study, we embark on the challenge of providing a proof of concept that translates the cognitive concept of *acquired embodiment* into a practical software framework. Our primary objective is to automatically model ACs by leveraging features extracted from art images that evoke these abstract notions. The choice of art images as a case study of this endeavor is driven by the allure of ever-expanding image collections, particularly the burgeoning art catalogues within the CH domain. These visual forms, encompassing paintings and photographs, illustrate, and circulate, concepts through what Barthes called an image's 'connotation' [32]. For example, seeing Artemisia Gentileschi's *Judith Beheading Holofernes* (1620 ca.), a human observer can detect objects such as a sword, but a comprehensive understanding of the painting would generally include an AC such as *violence*. In this context, the visual archives of CH often employ controlled thesauri like the Getty vocabularies and classification systems like Iconclass, replete with ready-made ACs that can be associated as subject matters for visual materials. However, the computational interpretation of these culturally coded visual elements is far from straightforward. Despite the remarkable progress in CV, the field falls short in the domain of AC image classification (see results of Chapter II.2). The predominant focus on image segmentation is ill-suited for abstract notions that lack the distinctive physical features present in many concrete objects.

This chapter's goal is to translate recent cognitive theories about concept representation into a software architecture that can automatically model ACs based on multimodal features. We hypothesize that a formal representation of ACs as multimodal frames can be automatically produced with a pipeline combining knowledge engineering, CV, and computational linguistics methods. In this work, we introduce our approach to generating these representations through the extraction and integration of features from images to KGs. Our approach focuses on the extraction, analysis, and integration of multimodal features (including depicted concrete objects, depicted actions, and color features) from images tagged with ACs. Taking the Tate Gallery collection as an empirical basis, we present a study of the extraction and integration of multimodal data. The contributions of this work can be summarised as follows:

- We define a conceptual model and present a novel, pattern-based ontology for AC representation, which allows for semantic characterization of multimodal features.
- We propose a novel approach for the extraction and integration of multimodal features of images that evoke certain ACs.
- We implement the proposed method on a corpus of art images from a well-

known catalogue of art images (the Tate Gallery collection).

- We discuss the results and provide all software, data sources, and results to allow the reproducibility of our experiment.

The remainder of this work is organized as follows. In Section III.1.2 we describe the Tate Gallery Collection's dataset, which was used as our experiments' input source. In Section III.1.3, we explain our approach, and in Section III.1.4 its experimental implementation. Section III.1.5 reports our experiments and results, while Section III.1.6 focuses on the discussion of the results and their limitations. We conclude with further directions in Section III.1.7

III.1.2 Input Source: The Tate Gallery

As an empirical basis for our study, we use the Tate Gallery's collection metadata of 70,000 artworks made available as a Github repository.[1] The Tate is an institution that houses the United Kingdom's national collection of British art, and international modern and contemporary art. Most of the collection is from the 1800s, and a considerable part of it is from after 1960. In 2013, the institution made its collection metadata available for about 70,000 artworks that it owns or jointly owns with the National Galleries of Scotland, through a GitHub repository. While the repository is no longer actively maintained, the Tate keeps it available as a useful tool for researchers and developers, and looks positively on creative remixing, visualization, and analysis of their collection metadata.[2]

The dataset contains complete records of most artists and artworks in the collection. It also includes image and thumbnail URLs, but it does not directly provide images, which still need to be accessed online. The dataset can be accessed in two ways: either through CSV files containing information about the artists and artworks, or in a series of thousands of text files containing all the records in JSON format. The JSON data is much richer than the CSV, storing a list of subjects associated with the record organized in a subject taxonomy.

It was precisely because of its subject taxonomy that we selected the Tate Gallery Collection as the first dataset to test our approach; the rich taxonomy includes both concrete concepts and ACs referring to non-physical objects ("vacuum cleaner," "shoe", "consumerism", "horror") as subject tags. As documented

[1] https://github.com/tategallery/collection. Access date: May 2021.

[2] See Eric Drass' "Tate Explorer" at http://shardcore.org/tatedata/ and Florian Kräutli's "The Tate Collection on Github" at http://research.kraeutli.com/index.php/2013/11/the-tate-collection-on-github/. Access dates: May 2021.

in an issue of their GitHub repository,[3] the Tate's subject taxonomy is a "bespoke taxonomy", originally developed alongside the digitization of Tate's collection as a means of enabling visitors to search artworks via subject. The design of the hierarchical structure and initial tagging of the bulk of the collection was expert-led, carried out by indexers with art history backgrounds and with the support of Tate's curatorial team, and in consultation with pre-existing systems such as Iconclass and COLLAGE–the now unavailable public-access system for the Guildhall Library and Art Gallery.

III.1.3 Approach

Our approach is based on the idea that an AC is a complex object whose definition can be formalized as a *description* [245]. Specifically, we assume that an AC can be described via a multimodal frame, whose meaning arises from an integration of sensory-perceptual and linguistic features of content, such as images, that evoke that AC. For example, the meaning of the AC *death* may be described via a multimodal frame integrating properties of linguistically co-occurring terms (e.g., "coffin", "gun", "blood") and sensory-perceptual properties of images depicting scenes that evoke that AC (e.g., the color black as a sensory-perceptual feature of a funeral scene, which evokes the AC "death") [101]. We developed a framework (summarized in Figure IV.1.2) to integrate multimodal features related to ACs into a scalable ontology-based KG.

ACs Candidate List Creation Determining a starting list of candidate concepts that can be reliably classified as referring to non-physical objects is necessary to begin this research. In our approach, the initial list is based on the conceptual taxonomy already in use by the Tate Collection to tag the content of visual artworks, which explicitly includes categories such as "universal concepts" and "social comment".

AC Definition by Multimodal Frames Our approach is based on the idea that an AC can be defined in a multimodal frame which describes and integrates complex linguistic and sensory-perceptual features. To represent this model formally, we designed the Multimodal Descriptions of Abstract Concepts (MUSCO) ontology, based on the Descriptions and Situations (DnS) ontology [129], which supports a first-order manipulation of theories and models. DnS was chosen as a core design pattern because it allows for the modeling of non-physical objects,

[3] https://github.com/tategallery/collection/issues/27. Access date: May 2021.

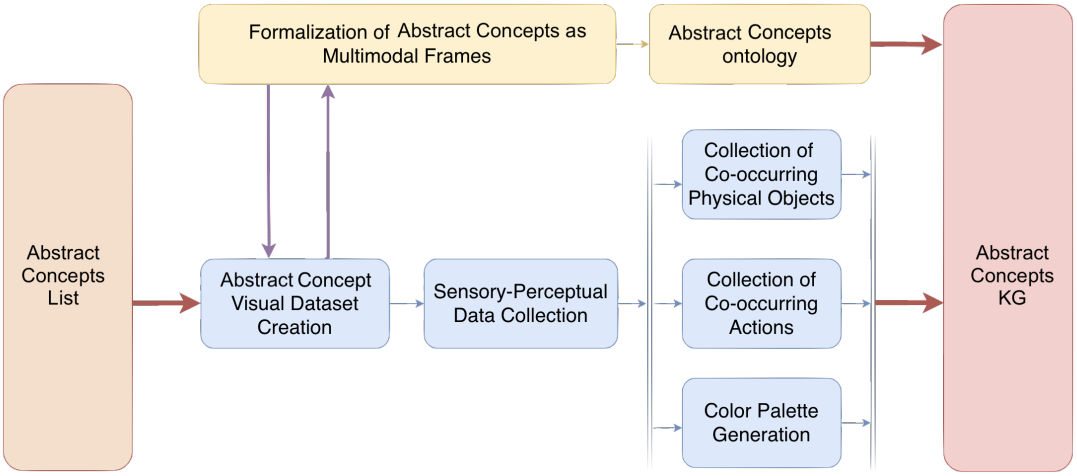


Figure III.1.1: The pipeline aims to populate a large-scale, ontology-based ACs KG that describes ACs with multimodal frames.

such as ACs, whose intended meaning results from statements, i.e. they arise in combination with other entities. Specifically, DnS axioms capture the notion of *situation* as a unitarian entity out of a State of Affairs (SoA), that is constituted by the entities and the relations among them, and a *description* as an entity that partly represents a (possibly formalized) theory that can be conceived by an agent. Influenced by the work in [361], in the MUSCO ontology we consider that the image annotation process is a situation (i.e. a context reified in the class `ImageAnnotationSituation`) that represents the state of affairs of all related, annotated data: actual multimedia data as well as metadata, and that needs to be described via an `ImageAnnotationDescription` (see Figure III.1.2).

We have engineered MUSCO's architecture in a deliberately modular way, such that an `ImageAnnotationDescription` can be modularized into subdescriptions that capture complex structures to be annotated. At this stage, we have identified three complex structures to be annotated in art images evoking certain ACs: dominant colors, depicted physical objects, and depicted actions. As such, in the MUSCO ontology we define the general `ImageAnnotationDescription`, which is satisfied by the general `ImageAnnotationSituation`, and which is composed by three more specific descriptions (`DominantColorDescription`, `raPORecognitionDescription`, and `ActionRecognitionDescription`) that define concepts and give meaning to data extracted in the context of each of the complex structures (see Figure III.1.3). Finally, the MUSCO Ontology defines the description class `SCMultiModalFrame`, which (a) defines a `SocialConcept`, (b) is used by a `ImageAnnotationSituation`, and (c) can be evoked by some `ImageObject`. The ontology also already allows for the conjunct expres-

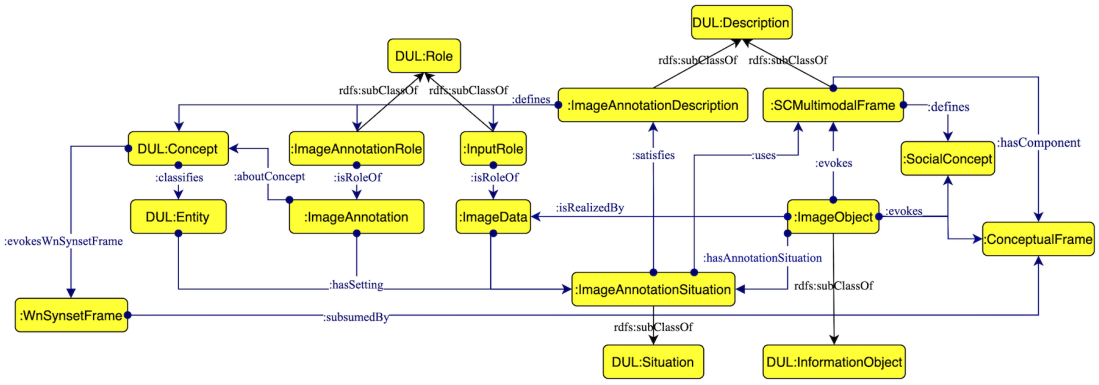


Figure III.1.2: The MUSCO Ontology is aligned to and reuses patterns from DOLCE+DnS Ultralite (DUL) foundational ontology in order to represent and give meaning to all data created during an image annotation process. All unprefixed classes belong to the namespace of the MUSCO ontology.

sion of data coming from linguistic and lexical resources through classes such as `WnSynsetFrame` and `ConceptualFrame`. Even though the current version of the MUSCO ontology is still under the process of revision and editing, in its current state, it already includes the entities, resources, and relationships necessary to integrate the experimental outputs presented in this work.

Abstract Concept Visual Dataset Creation For each AC, a corpus of (art) images that have previously been explicitly tagged with that AC is created. This corpus is produced by performing surveys of existing image catalogues which include that AC in their tagging scheme.

Sensory-Perceptual Data Collection Our approach is based on the idea that sensory-perceptual features of ACs can be extracted from images that evoke those ACs. For each AC, based on its corpus of images, we extract a set of features including: labels for concrete objects, actions depicted, and dominant color palettes. The specific technique for extracting concrete object labels from the images depends on the previously provided information from that image. If the images have already been tagged with labels of concrete objects, these are collected for further analysis. If they have not, an object recognition task especially attuned to art images is performed based on [90] to recognize physical objects in the image and gather labels for such objects. A similar approach is taken for extracting labels for depicted actions (e.g., "sitting", "standing") and/or relationships (e.g., "holding hands", "hugging"). If tags for actions or relationships have already been attached

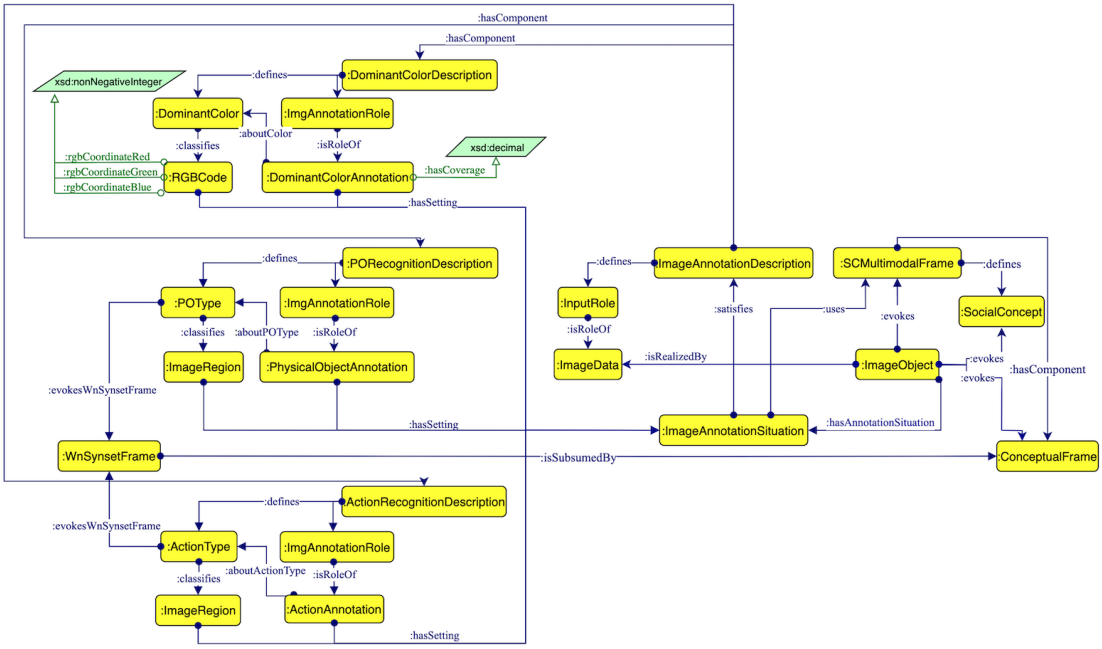


Figure III.1.3: The MUSCO Ontology's reuse of the DnS pattern is modular: defining a general description for the image annotation situation, composed of simpler, more specific descriptions which give meaning to specific annotation structures and results. All classes in the figure belong to the namespace of the MUSCO ontology defined in this work.

to the images in the corpus, these are used. If not, we perform relation and action detection on the images following the most recent relation modeling techniques after they have been trained specifically to process art images, following the approach by [190]. Finally, color analysis is performed on the images following the method provided by the extcolors Python package.[4] This technique groups colors based on visual similarities using the CIE76 formula, and outputs both an image (for visual representation) and a text result with the usage (in number and percentage of pixels) of the top five dominant (most used) colors in that image.

III.1.4 Experimental Set-Up

III.1.4.1 Experiments

Input Data Preprocessing. To apply our method to the Tate Gallery Collection dataset, it was first necessary to reconstruct the hierarchy of Tate's subject

[4] https://pypi.org/project/extcolors/. Access date: May 2021.

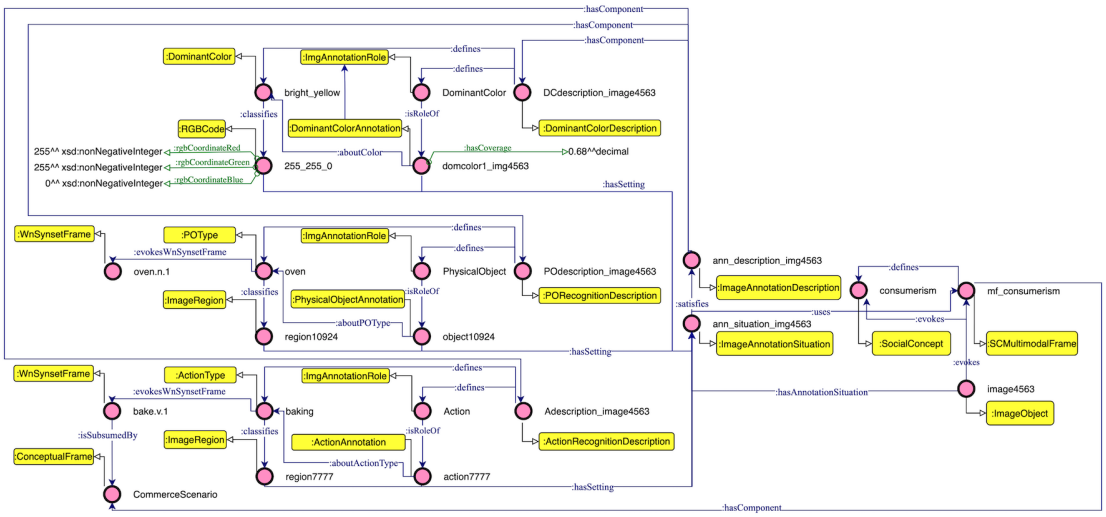


Figure III.1.4: Example use of the MUSCO ontology to formalize multimodal features extracted from one image (dominant color *bright yellow*, depiction of the physical object *oven*, and depiction of the action *baking*), the image's evocation of the AC "consumerism", and the concept's description as a multimodal frame. All arrows with white arrowheads stand for the relation `rdf:type`.

taxonomy in a way that allowed its eventual integration into an ACs KG using the MUSCO ontology. For this task, we performed a survey of the taxonomy based on metadata accessed in March of 2021, finding that (1) it has three hierarchical levels, going from broadest to narrowest categories, and that (2) it includes hierarchical relationships between 2409 subject tags. These 2409 subject tags were the ones used to reconstruct and then formalize Tate's tagging hierarchy. To integrate the taxonomy and its subjects into our KG, we extended the MUSCO ontology by reusing concepts defined in the Simple Knowledge Organization System (SKOS) data model[5] (see Figure III.1.5). With this extension, we were able to represent hierarchical relations between subject tags, specifically with the property `skos:broader`. We implemented a Python script to transform and serialize the taxonomy from JSON to Turtle (.ttl) format[6]. All resources (MUSCO ontology, .ttl file, Python functions) are available in the Github repository.[7]

Abstract Concepts Candidate List Creation. Visualizations of the Tate's subject taxonomy as graphs (also available in the GitHub repository) were per-

[5] `https://www.w3.org/TR/skos-reference/#schemes`. Access date: May 2021.
[6] `https://github.com/tategallery/collection/tree/master/processed/subjects`. Access date: May 2021.
[7] `https://github.com/delfimpandiani/musco`. Access date: May 2021.

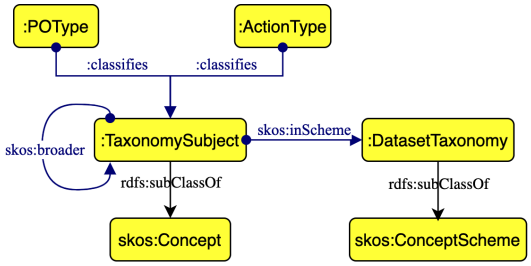


Figure III.1.5: Addition to the MUSCO ontology model to formalize the use of concept schemes coming from collections' or other datasets' taxonomies, such as Tate's subject taxonomy. All classes with no explicit namespace belong to the namespace of the MUSCO ontology.

formed with the Graphviz[8] package, in order to ease manual investigation of its coverage and identification of areas where ACs may be more pervasive. Three areas of interest emerged: first, the level 0 category "emotions, concepts and ideas" (specifically its level 1 children "universal concepts" and "emotions and human qualities"); second, the level 1 category "social comment" (child of level 0 category "society"), and third, the level 0 category "religion and belief" (see Figure III.1.6). A total of 166 "narrow" [level 2] ACs were manually selected from these categories (80 from "emotions, concepts, and ideas", 67 from "society", and 19 from "religion and belief"). These concepts' parent [level 1] and grandparent [level 0] tags were excluded from subsequent analysis for two reasons. First, compared to their narrow children tags (e.g., "fear", "education"), many of the broader tags actually refer to multiple ACs at once (e.g., "emotions, concepts and ideas", "education, science and learning"). Secondly, a manual investigation of the Tate artworks' metadata showed that artworks seem to be explicitly tagged with level 2 tags, and level 1 and level 0 tags are only included by being higher in the hierarchy. That is, there seem to be no Tate artworks that are tagged only with the broader level 1 or 0 tags. The complete list of ACs and their parents is available in Appendix 1.

Concept-Artwork Matching. For each of the 166 ACs, the number of artworks tagged with that AC was extracted and recorded. Additionally, data about each artwork, including its bibliographic information (name, id, artist, date, etc.) was stored.

Co-occurring Subject Tags. For each AC, an investigation of the metadata of the artworks tagged with it was performed, collecting other subject tags used to index the content of those artworks. In this way, for each AC, we created a

[8] https://pypi.org/project/graphviz/. Access date: May 2021.

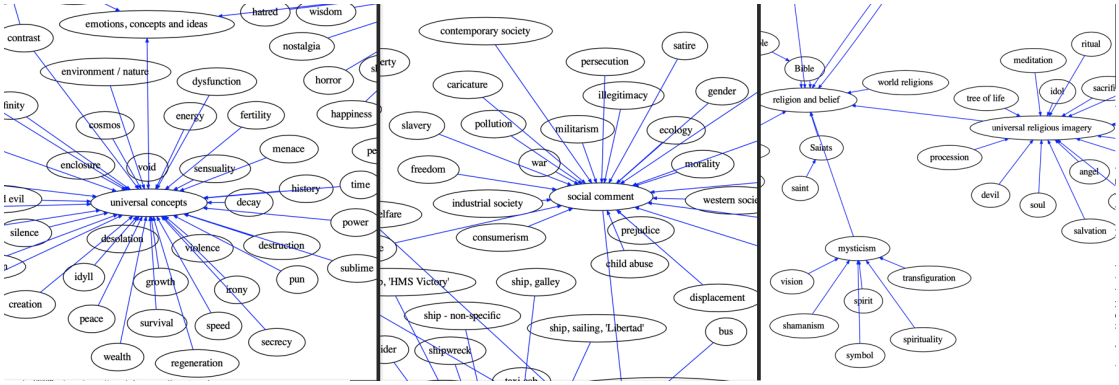


Figure III.1.6: Three main areas of interest for the identification of ACs within the Tate subject taxonomy. Social concepts such as "destruction", "peace", "wealth", "courage" (surrounding "emotions, concepts and ideas"), "consumerism", "freedom", "slavery", "nationalism", "ecology" (surrounding "society"), "magic", "enchantment", "worship", "blessing" (surrounding "religion and belief"), among others, were identified in these areas.

dictionary holding all co-occurring subject tags and the frequency of co-occurrence. This allowed for the collection of object and action tags without having to resort to object or human-object interaction recognition techniques.

Co-occurring Physical Objects. Symbol grounding and *acquired embodiment* is expected to occur with labels referring to physical referents. Therefore, we wanted to identify all co-occurring tags that specifically refer to physical objects. The Tate's subject taxonomy allowed for a quick solution to this goal, as its hierarchical organization includes the top [level 0] concept "objects", which based on a manual investigation seem to refer to physical objects. We extracted all tags under this category. A manual inspection of the initial results showed that certain physical entities, such as "woman", or "tree" were not being extracted from artworks that clearly depicted these entities (and were tagged with them). From this observation, we noted that additional labels referring to physical objects were under the categories "people" and "nature", so we also extracted all tags whose parent tag was "children", "adults", or "nude", as well as those tags whose parent or grandparent tag was "nature". While it was not possible to confirm that absolutely all terms referring to physical objects were extracted, after a manual inspection of the Tate's subject taxonomy, it was concluded that a large majority of them was indeed extracted. We then performed statistical analyses to obtain the number and frequency of co-occurring physical objects for each AC, as well as the top ten most frequent physical objects co-occurring with each AC. We also calculated the averages and medians of these measures by taking into account all

ACs and their co-occurring physical objects.

Co-occurring Physical Actions. A similar process to the one just described for labels referring to physical objects was performed for what, within the Tate's subject taxonomy, are "action" labels. The identification of tags that refer to actions was completed by identifying tags under the categories "actions:postures and motions", "actions:processes and functions", or "actions:expressive", and "animals:actions". We extracted all action tags co-occurring with each of the ACs, and then performed statistical analyses as described in the previous section.

Dominant Color Visual Analyses. We then performed color analyses related to two ACs selected as case studies: "consumerism" and "horror". The hypothesis supporting this choice was that they should show distinctive color profiles. For each concept, the color analysis was performed with the extcolors Python package mentioned above, on 30 randomly selected images of artworks (specifically, of paintings and prints) that had been explicitly tagged with the concept. For each image, a color palette was created by extracting the RGB coordinates and occurrence rates (in both number and percentage of pixels) of the five colors with the highest occurrence rate in that image. The criterion for this choice was that identifying the colors with the highest occurrence in an image is a proxy for identifying the most pervasive/visible colors in an image, which may be a relevant feature humans use to judge whether images evoke certain concepts. To represent this idea more intuitively, a further analysis was completed to generate images of the *proportional* palettes, so as to represent the most common colors present in each of the artworks more intuitively. This final task was completed with the MulticolorEngine by TinEye,[9] but we are developing code for the automation of this procedure.

Data Integration. We automatically incorporate the extracted data, including the co-occurrence patterns and visual features–into the ACs KG via the MUSCO ontology.

III.1.5 Results

Based on the metadata accessed in March of 2021, the Tate's subject taxonomy was found to be divided into three levels: [level 0] top concepts, representing the most general categories; [level 1] slightly narrower concepts, children of level 0 concepts; and [level 2] narrowest concepts, children of level 1 concepts and grandchildren of level 0 concepts. Out of the 2409 subject tags in the taxonomy, 16 are level 0 concepts, 142 are level 1 concepts, and 2251 are level 3 concepts. The 166 ACs

[9] https://labs.tineye.com/color/. Access date: May 2021.

selected as initial candidates are all grandchildren of 3 out of the 16 top concepts, and children of 15 out of the 142 middle categories available (see Table III.1.1).

Concept-Artwork Matching. The number of Tate artworks explicitly tagged with each of the 166 chosen ACs ranged from 368 ("death") to 1 ("paranoia"), with an average of 48 matches and a median of 27 matches. The two case studies were in the top 20% of ACs ranked on artwork matches ("consumerism" with 71 artworks, and "horror" with 146 artworks) (see Table III.1.2).

Co-occurring Physical Objects and Actions. The number of co-occurring tags for each of the 166 chosen ACs ranged widely, from 1506 ("death") to 7 ("paranoia"), with an average of 311 co-occurring subjects and a median of 262. Further analyses were performed to identify physical objects and action tags from the co-occurring subject tags for each AC. The number of co-occurring *physical object* tags for each of the chosen ACs ranged from 288 ("death") to 6 ("infinity"), with an average of 69 co-occurring physical objects and a median of 55 physical objects. Table III.1.3 shows the top ten most frequent co-occurring physical objects for four ACs. The number of co-occurring *action* tags for the 166 ACs was decisively smaller, ranging from 38 ("death") to 0 ("void"), with an average of 11 co-occurring actions and a median of 8 actions. Table III.1.4 shows the top ten most frequent co-occurring actions for four ACs. The average frequencies of co-occurrence for physical objects and for actions with each AC are also presented in Table III.1.2. Finally, Figure III.1.7 includes more intuitive visual representations for most co-occurring physical objects (top) and actions (bottom) for the two case study ACs.

Level	**# Tags**	**Concrete Tags**	**# ACs**	**Soc. Concept Tags**
0	16	"objects", "nature", "people"	3	"society"
1	142	"weapons", "trees", "adults"	15	"social comment"
2	2251	"missile", "oak", "old man"	166	"consumerism"

Table III.1.1: Distribution of Tate subject tags based on its three hierarchical levels, going from broadest (0) to narrowest (2). Each level includes concrete concepts referring to physical objects and ACs referring to non-physical objects.

Dominant Color Visual Analyses. Figure III.1.8 presents visual representations (as proportional color palettes) of the color palette analyses performed on 30 randomly chosen images of paintings and prints for each of the two case study ACs ("consumerism" and "horror").

AC	**# Matches**	**# CO-O**	**Freq Top CO-O**	**# CO-A**	**Freq Top CO-A**
death	368	288	71.1	38	14.5
horror	146	138	33.5	30	6.0
consumerism	71	129	16.4	7	2.4
paranoia	1	15	1.7	5	1.2
Average	*48*	*69*	*9.7*	*11*	*2.8*
Median	*27*	*55*	*5.3*	*8*	*1.6*

Table III.1.2: Number of artwork matches, co-occurring physical objects (CO-O), and co-occurring actions (CO-A), along with the average frequency of co-occurrence for the top ten most frequently co-occurring physical objects (Freq Top CO-O) and top ten actions (Freq Top CO-A) for four ACs: "death" (with the highest number of matches), "paranoia" (lowest number of matches), "consumerism" and "horror" (case studies). Average and median values calculated from all 166 ACs are also provided.

Abstract Concept	**Top 10 Physical Objects**
death	'man', 'woman', 'religious and ceremonial', 'clothing', 'furnishings', 'male', 'weapons', 'female', 'fine arts and music', 'sea'
horror	'man', 'clothing', 'woman', 'uniform', 'animal/human', 'reading, writing, printed matter', 'male', 'fine arts and music', 'monster', 'medical'
consumerism	'woman', 'reading, writing, printed matter', 'clothing', 'furnishings', 'food and drink', 'domestic', 'electrical appliances', 'kitchen', 'tools and machinery', 'product packaging'
paranoia	'man', 'clothing', 'woman', 'figure', 'male', 'furnishings', 'curtain', 'jacket', 'jumper', 'suit'

Table III.1.3: Top ten most frequent co-occurring physical objects with the AC with the most matched artworks ("death"), for the AC with the least matched artworks ("paranoia"), and for the two case studies ("consumerism" and "horror").

Abstract Concept	**Top 10 Actions**
death	'standing', 'lying down', 'reclining', 'supporting', 'embracing', 'sitting', 'flying', 'kneeling', 'carrying', 'sleeping'
horror	'standing', 'sitting', 'recoiling', 'watching', 'lying down', 'screaming', 'carrying', 'hand/hands raised', 'embracing', 'fleeing'
consumerism	'smiling', 'sitting', 'crouching', 'standing', 'reclining', 'lying down', 'talking'
paranoia	'sitting', 'crouching', 'reclining', 'standing', 'walking'

Table III.1.4: Top ten most frequent co-occurring actions with the AC with the most matched artworks ("death"), for the AC with the least matched artworks ("paranoia"), and for the two case studies ("consumerism" and "horror").

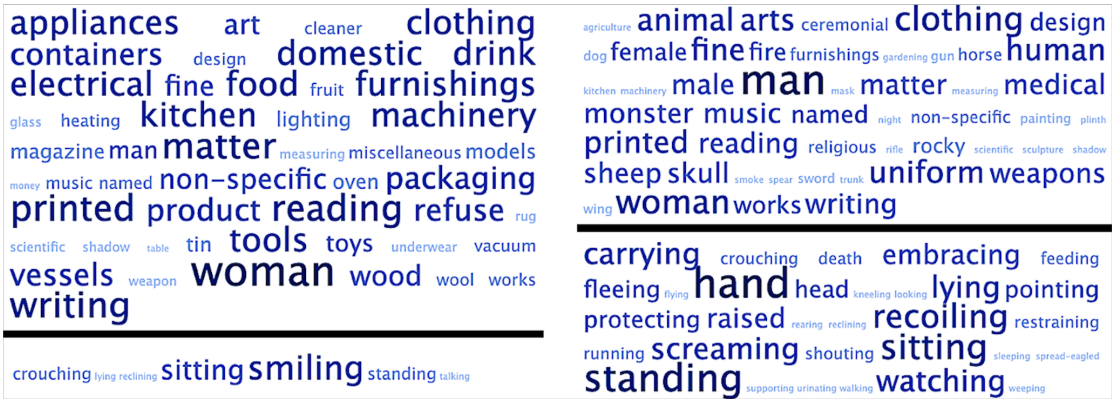
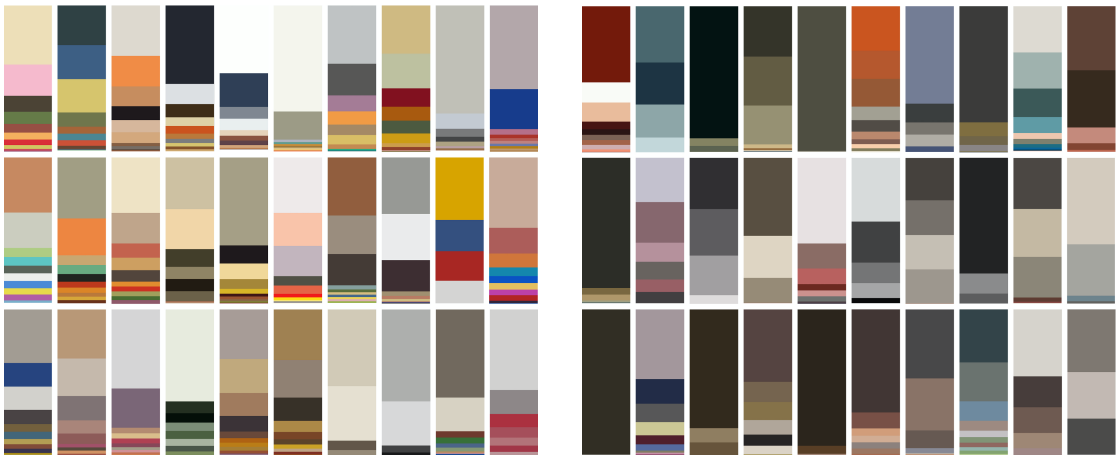


Figure III.1.7: Wordclouds for the top 50 co-occurring object (top) and action (bottom) tags for all artworks tagged with "consumerism" (left) and with "horror" (right). Larger words more frequently co-occurred with the AC of interest.



(a) Consumerism (b) Horror

Figure III.1.8: Proportional palettes of 30 paintings and print images tagged with "consumerism" (left) and with "horror" (right).

III.1.6 Discussion

In our reconstruction and examination of the Tate's subject taxonomy, we found that ACs referring to *non-physical* objects were concentrated in three major areas of the taxonomy ("emotions, concepts, and ideas", "society", and "religion and belief"). Gathered from the tagging taxonomy of a well-known collection of art images, the initial list of ACs we present stands as empirical proof of the use of ACs for tagging the content of visual material. With this work, we begin the creation of a corpus of (art) images tagged with ACs, currently available in the GitHub repository as a dictionary with ACs as keys, and lists of image URLs as values.

Our experiments on the co-occurrence of physical objects and actions further show that it is possible to develop computational techniques that mirror the idea of symbol grounding and *acquired embodiment*, i.e., it is possible to identify perceptual features of concrete objects and actions that co-occur with ACs, at least in the context of art images. While there seems to be some regularity in the results, the low frequencies of co-occurrence suggest that further research is needed to understand which of these objects and actions, if any, have a substantial effect on the evocation of an AC. A manual examination of the top ten most frequent concepts co-occurring with ACs (some of them presented in Table III.1.3 and Table III.1.4) suggest some regularity in the physical objects and actions that most frequently co-occur with certain ACs in these art images. Most of these co-occurrences seem to agree with intuition (i.e., "consumerism" co-occurring with "clothing", "food and drink", and "product packaging"; "horror" co-occurring with "monster", "recoiling" and "screaming"). The results of the color analyses, visible in the proportional color palettes presented in Figure III.1.8, also strongly suggest a certain degree of regularity in the color features of Tate art images that evoke certain ACs. As with the top co-occurrence results, at first sight, the color palettes seem to agree with intuition (i.e, "consumerism" shows a greater luminosity and variety of bright colors, as in the aisles of a supermarket; "horror" shows dark and less varied colors and tones, as in typical scenes from horror movies).

Some limitations of this work should be noted, such as the fact that the images used in the experiments were not distinguished by the type of artwork medium (painting, print, drawing), which may affect the type and frequency of features that were extracted from the images. It is also important to note that the regularity observed from the experimental results may be limited only to the Tate's collection and not generalizable to other art image catalogues or collections. That is, while the Tate dataset provided us with a clean, annotated corpus in which ACs are explicitly used as tags, as a curated dataset it encompasses a limited geographic, historical, and cultural perspective. Additionally, the fact that it is expert-tagged might result in a biased interpretation and classification of the artworks, which

might differ from the interpretations of a larger, more diverse group of viewers. Given ACs' ambiguous meanings, further work is needed which takes into account various possible interpretations of the same artwork by different observers.

Further directions for this work are multiple. The generated KG may be additionally populated with other methods, for example by automatically generating descriptive paragraphs from the art images, and then performing knowledge extraction on the natural language paragraphs, as well as by including additional sensory data, such as sound or smell data, that may evoke certain ACs. We can also improve our approach by refining the initial ACs list through alignment with the latest cognitive science research, as well as through user-based studies. Additionally, disambiguating the terms, expanding the terminology by leveraging lexical resources such as WordNet, VerbNet, and FrameNet, and studying the terms' distributional linguistic features in a textual corpus are next steps that could lead to substantial improvements. The visual analyses can be further refined by distinguishing artwork medium types, and by extracting contrast measures, common shapes, repetition, and other visual patterns. A particularly interesting research direction would be to include facial recognition analyses in our pipeline, which may allow the identification of emotions expressed by depicted subjects.

III.1.7 Conclusions

Our approach models ACs as complex objects, which can be described with multimodal frames that integrate multisensory information. To represent this model formally, we designed the pattern-based ontology Multimodal Descriptions of Abstract Concepts (MUSCO), which allows for the conjunct semantic characterization of multimodal features. Our approach also proposes a pipeline for automatically extracting and integrating features of images that evoke ACs. We show its potential by testing it on a corpora of art images from a well-known collection. Our experimental results point towards some regularity in certain sensory features of images tagged with specific ACs, and open space for further research to evaluate the proposed approach focusing on datasets with different characteristics. More than anything, our results serve as a proof of concept and open up new lines of future research. The automatic population of a KG with the extracted data is the natural next step of this work, potentially through mapping-based knowledge extraction techniques. Our method is impulsed by the intuition that a KG containing multimodal data for AC description can eventually serve as input to a learning model to automatically detect ACs in images. Overall, the experiments performed and results presented in this work serve as a proof of concept that extracting, integrating, and coincidentally exploiting multimodal data related to ACs is a promising direction for future research.

Chapter III.2

Perceptual Semantics: Shallower Waters, Clearer Insights

Summary In this chapter, we delve into our *perceptual semantics* paradigm, shifting away from conventional reliance on end-to-end deep visual features. Instead, we explore the use deep models within more concrete, 'shallower' strata of the semantic pyramid, with the aim of obtaining more interpretable image representations. Our approach entails a shift from the direct application of deep learning to bridge raw pixels to ACs, instead favoring a feature engineering strategy that harnesses the profound capabilities of DL models to capture the essence of more tangible semantics. Building upon the insights gained from earlier chapters (Chapters II.2 and III.1), we employ a suite of off-the-shelf deep learning detectors to autonomously extract perceptual semantics inspired by cognitive processes, ensuring that the DL black box remains grounded at more concrete levels. These extracted semantics serve as features for image representation, denoted as I_{PS}, enabling us to strike a middle ground that enhances clarity and interpretability. Subsequently, we explore the use of these cognitive-inspired image representations to enhance performance and explainability in AC image classification. These more interpretable image representations serve as input for traditional machine learning models, as opposed to deep learning ones, to conduct image classification. Our primary objective throughout this process remains constant - maximizing class probabilities based on the transformed image representation and model parameters. We achieve this using classical machine learning models, such as Naïve Bayes, Random Forest, and others. The results obtained from this methodology demonstrate that by prioritizing feature engineering and traditional machine learning methods over DL, we not only maintain similar performance levels to Convolutional Neural Networks (CNNs) but also significantly enhance the interpretability of our system for AC image classification.

III.2.1 Introduction

The ARTstract dataset, introduced in Chapter II.1, is a curated collection of cultural images used as a resource to train models on the challenging goal of classifying images based on abstract concepts (ACs). As such, it contains images labeled with ACs such as *comfort*, *danger*, *death*, *fitness*, *freedom*, *power*, and *safety*. The ARTstract image dataset aligns with the traditional deep learning (DL) paradigm, employing solitary, decontextualized target labels. However, our earlier research in Chapter I.2 emphasized the crucial role of explicit semantics in human high-level understanding of visual content, in which the detection of multiple layers of meaning plays a pivotal role in assigning abstract interpretations [33]. This multi-faceted approach stands in contrast to the one-dimensional nature of the ARTstract image dataset, where each image is assigned a single overarching AC label, such as *comfort*. Consequently, the dataset lacks labels representing more granular and concrete layers of semantics, including objects, actions, colors, emotions, and other specific details (e.g., *couch*, *sitting*, *brown*, and so on) that humans rely on to inform their decisions.

At the same time, Chapter III.1 has provided evidence of the potential of these visual data descriptors to bridge the semantic gap between raw pixels and ACs. While Chapter II.2 underscores the challenges faced by DL models in AC image classification, particularly in providing explanations for their predictions, it also highlights their proficiency in extracting concrete image features. This proficiency is evident from significant performance differences in image classification across datasets like CIFAR and ARTstract (see Section II.2.5 of Chapter II.2). These distinct strengths and challenges of the DL paradigm, combined with the potential of cognitive-inspired perceptual semantics, raise a fundamental question:

> *Can cognitive-inspired perceptual features be effectively leveraged and employed to enhance image representations for improved performance and explainability in the context of AC image classification?*

Our hypothesis posits that perceptual semantic (PS) features inspired by the idea of acquired embodiment can play a pivotal role in crafting interpretable image representations that are compatible with statistical methods from classical machine learning (ML). The core idea revolves around the engineering, extraction, and integration of image features to improve the performance and interpretability of AC image classification This approach emphasizes the strategic containment of deep learning's inherent black box, limiting its role to associating more concrete perceptual semantics with images and making these associations explicitly available to the model and to the user. This strategic containment opens up possibilities for the systematic identification of problematic data sources and offers avenues for addressing issues pertaining to bias and subjectivity in image annotation processes.

This chapter accomplishes the following:

- **Perceptual Semantics (PS) Extraction**: It presents a framework for fully automatic extraction of a wide range of perceptual semantics from ARTstract images.
- **Perceptual Semantics (PS) Image Representation**: The relevance of each extracted perceptual feature in predicting abstract concepts is quantified using information theory, involving the calculation of feature entropy conditioned on the evoked abstract concept. These findings are then visually represented, both in a general context and concept-wise, through relevant graphs.
- **PS-Based AC Image Classification**: To model the contextual dependencies among these features, we employ a series of classical ML approaches to learn the joint probabilities. These are then leveraged for AC image classification, predicting the class of the held-out portion of the dataset. Performance and explainability are explored and discussed.

In this chapter, we begin with Section III.2.2 in which we explain our idea of using perceptual-semantics based image representations for more interpretable AC image classification. Then, in Section III.2.3, we discuss multiple aspects of our approach: the selection process for concrete labels and methodology for automatically extracting perceptual knowledge from over 14,000 ARTstract images (subsection III.2.3.1), the analysis of ARTstract's perceptual semantics, identifying common and significant elements for each target AC cluster (subsection III.2.3.2), and our novel approach to image representation and AC image classification, emphasizing the use of perceptual semantics and explainable, classical machine learning models (subsection III.2.3.3). In Section III.2.4, we discuss our results, which we discuss in depth in Section III.2.5. We provide valuable insights into the advantages of our method in terms of interpretability, while outlining potential directions for future research.

III.2.2 Idea: Perceptual Semantics (PS)

Part II highlighted the robustness of deep models in handling more concrete visual tasks. Thus we focus on the extraction of cognitive-based concrete intermediary features, referred to as *perceptual semantics* (PS), for the purpose of image representation. To accomplish this, we introduce a function f_{PS} designed to process the raw image representation I_{RAW} and transform it into a novel representation, I_{PS}, which takes the form of a vector within $\mathbb{R}^N$ (see Figure III.2.1). Specifically, f_{PS} is constructed as a composite function consisting of M individual detectors: $f_{PS} = [f_{PS_1}, f_{PS_1}, \ldots, f_{PS_M},]$ (see Figure III.2.2). Each of these detectors yields N_M labels. Subsequently, these perceptual semantic labels are aggregated to yield the total dimensionality:

$$f_{PS_1} : I_{RAW} \rightarrow \mathbb{R}^{N_1} \tag{III.2.1}$$
$$f_{PS_2} : I_{RAW} \rightarrow \mathbb{R}^{N_2} \tag{III.2.2}$$
$$\vdots \tag{III.2.3}$$
$$f_{PS_M} : I_{RAW} \rightarrow \mathbb{R}^{N_M} \tag{III.2.4}$$

$$N = N_1 + N_2 + \cdots + N_M \tag{III.2.5}$$

The final output of the f_{PS} function transforms the raw image (I_{RAW}) into a vector space with a dimensionality of $\mathbb{R}^N$:

$$f_{PS}(I_{\text{RAW}}) = I_{\text{PS}} \subseteq \mathbb{R}^N \tag{III.2.6}$$

We can utilize this new I_{PS} representation in our problem formulation and train a classical ML predictor, such as Naive Bayes, to make inferences:

$$\hat{y} = \arg\max(p(y_i|I_{\text{PS}}, \theta)) \tag{III.2.7}$$

The formulation of the problem in a classical ML way with this image representation is significantly more explainable because we can identify the top features, which are specific perceptual semantics, denoted as PS_{f_n}, that contribute to the highest probability of a particular class.

$$p(y_i|I_{\text{PS}}, \theta) = p(y_i|I_{\text{PS}_{f_0}}, \theta) + p(y_i|I_{\text{PS}_{f_1}}, \theta) + \cdots + p(y_i|I_{\text{PS}_{f_N}}, \theta) \tag{III.2.8}$$
$$p(y_i|I_{\text{PS}}, \theta) = \sum_{n=0}^{N} p(y_i|I_{\text{PS}_{f_n}}, \theta) \tag{III.2.9}$$

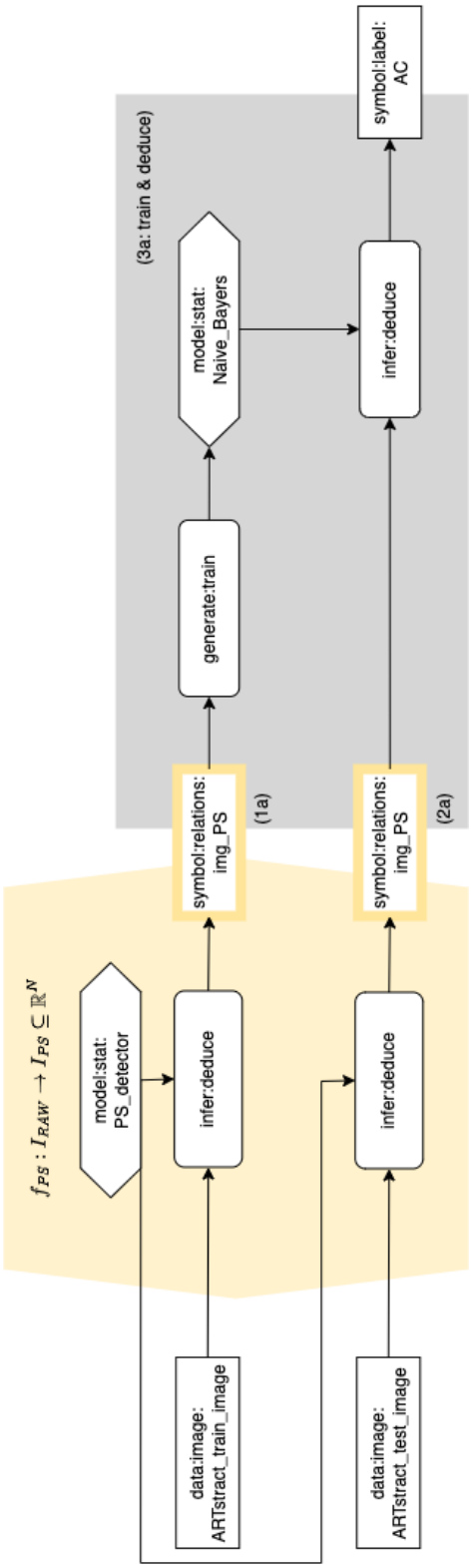


Figure III.2.1: Architecture of the perceptual semantics approach to AC image classification.

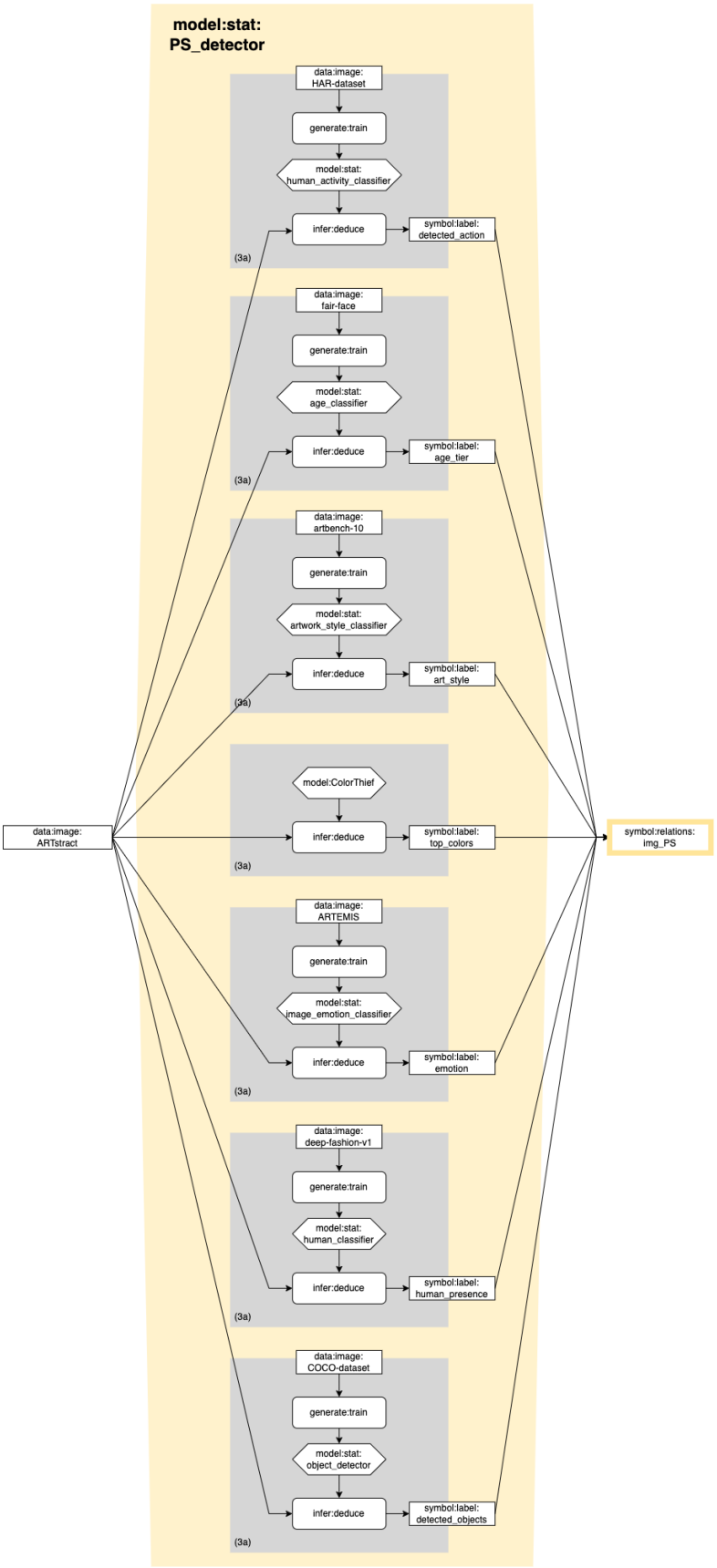


Figure III.2.2: Architecture followed to extract perceptual semantics from each image in the ARTstract dataset using automatic annotators, primarily pre-trained deep learning models.

III.2.3 Approach

PS Unit	Artificial Annotator	Ann. Type	Model Backbone	Pretrained on Dataset
Action	DunnBC22/vit-base-patch16-224-in21k_Human_Activity_Recognition	DL	Visual Transformer	har-dataset
Age Tier	nateraw/vit-age-classifier	DL	Visual Transformer	fair-face
Art Style	oschamp/vit-artworkclassifier	DL	Visual Transformer	artbench-10
Top Colors	ColorThief	Stat	Color-Thief	*N/A*
Emotion	artemis_image-emotion-classifier	DL	ResNet CNN	Artemis
Human Presence	adhamelarabawy/fashion_human_classifier	DL	Visual Transformer	deep-fashion-v1
Image Caption	Salesforce/blip-image-captioning-large	DL	Visual Transformer	COCO-dataset
Detected Objects	facebook/detr-resnet-101	DL	Detection Transformer	COCO-dataset

Table III.2.1: Perceptual Semantic (PS) units and the artificial annotators chosen to detect them, along with information about their types, model backbones, and pretraining datasets. DL: Deep Learning, Stat: Heuristics-based.

III.2.3.1 Perceptual Semantics Selection and Detection

Drawing from insights from our examination of high-level visual understanding in CV (Chapter I.2) and the potential of acquired embodiment to mitigate the semantic gap (Chapter III.1), we carefully selected specific perceptual semantics to extract for each image in the ARTstract dataset. These included the most likely depicted action, age tier, art style, top colors, evoked emotion, presence of humans, depicted objects, and an automatically generated image caption (see Fig. III.2.3). Actions portrayed in images can convey symbolic meaning and contribute to the overall narrative or theme of the artwork, as abstract concepts, like actions, rely on semantic similarity relations [92, 91]. Analyzing the age tier of depicted characters or subjects can provide contextual insights into the historical or cultural

setting of the artwork, while examining the art style can facilitate identifying artistic movements, influences, and themes present in the artwork. Colors are pivotal in conveying mood, atmosphere, and symbolism in art; thus, identifying the dominant colors in an image can offer insights into the emotional tone, thematic elements, and aesthetic preferences of the artist. The inclusion of evoked emotion aligns with research suggesting the significant role emotions play in AC modeling and perception [205, 370]. Similarly, considering age tier and the presence of humans underscores the importance of contextual elements in art analysis, as these factors can influence the narrative, mood, and societal themes depicted in the artwork. The presence or absence of human figures can impact interpretation and emotional resonance, serving as focal points or narrative elements signifying various themes, relationships, or societal issues. Depicted objects can carry symbolic or representational meaning, enriching the viewer's understanding of the artwork's themes and messages.

To identify suitable models, we conducted a manual investigation, utilizing the Hugging Face[1] interface to identify the most downloaded and highly-rated off-the-shelf detection models specifically trained or finetuned for one of these tasks. We subsequently evaluated their outputs through a manual qualitative inspection, to ensure that the identified semantics were coherent. When necessary, we explored alternatives, again assessed their coherence, and incorporated more culturally or artistically specific detectors, such as emotion detectors. Table III.2.1 provides an overview of the chosen perceptual semantic features, the selected artificial annotators, the architectural backbones of the models (if applicable), and the datasets on which the artificial annotators were pre-trained (if applicable).

Importantly, while certain datasets that ARTstract was curated from did include some of these perceptual semantics (e.g., object and action tags in the Tate Gallery dataset, or objects in ADVISE), we deliberately decided not to rely on them. Instead, we followed the same process for all images within the ARTstract dataset. This decision was motivated by our desire to ensure comparability across all semantics and to explore the extent to which semantic data processing can be automated solely from the visual content of ARTstract (raw pixels and tagged AC clusters). We aimed to use detectors widely employed within the computer vision community, as represented by the Hugging Face repository, to better mirror how powerful and commonly adopted computer vision tools can interpret images at different levels of concreteness, without relying extensively on background knowledge. Consequently, this approach enables new test images or unseen pictures to undergo the same automated process, without necessitating human-annotated ground truths. More details about each detection can be found below, including if and how each perceptual semantic label was assigned a ConceptNet node.

[1] https://huggingface.co/models. Access date: July 2023.

Action Detection For identifying actions depicted in images, a Visual Transformer model was employed.[2] This model, fine-tuned for human activity recognition, was pre-trained on the Human Activity Recognition (HAR) dataset, with possible targets including *calling*, *clapping*, *cycling*, *dancing*, *drinking*, *eating*, *fighting*, *hugging*, *laughing*, *listening to music*, *running*, *sitting*, *sleeping*, *texting*, and *using laptop*. We retained the action with the highest probability and assigned each to the ConceptNet concept with the same word (e.g., the *running* tag was assigned the ConceptNet node `conceptnet:running`. The only exception was the tag *using laptop*, which was assigned the ConceptNet node `conceptnet:laptop`.

Age Tier Detection To determine the age tier of individuals in the images, a Visual Transformer model was employed.[3] This model was fine-tuned for age classification, using age tier categories ranging from *0-2* to *70+*, as defined in the age tier mapping. There are a total of 8 targets, each representing a different age group. The model was pre-trained on the Fair Face dataset and was capable of categorizing individuals into their respective age tiers. The age tier with the highest probability was retained for each image. We mapped each of the numerical age tiers to specific ConceptNet nodes based on their age ranges. For example, age tier '0-2' was assigned to the ConceptNet node `conceptnet:toddlerhood`, while age tier '3-9' was linked to `conceptnet:childhood`. This mapping continued for all age tiers, associating each age group with the corresponding ConceptNet representation, allowing us to enrich our age-related annotations with meaningful semantic context.

Art Style Detection To identify the artistic styles depicted in images, we employed a Vision Transformer (ViT) model.[4] This model was specifically fine-tuned for art style classification and was capable of recognizing various artistic styles, such as *Art Nouveau*, *Baroque*, *Expressionism*, *Impressionism*, *Post-Impressionism*, *Realism*, *Renaissance*, *Romanticism*, *Surrealism*, and *Ukiyo-e*. The model was pre-trained on the ArtBench-10 dataset [225] to provide accurate classifications. For each image, we retained the art style with the highest probability. We retained the action with the highest probability and assigned each to the ConceptNet concept with the same word (e.g., the *Post-Impressionism* tag was assigned the ConceptNet node `conceptnet:post_impressionism`).

[2] `https://huggingface.co/DunnBC22/vit-base-patch16-224-in21k_Human_Activity_Recognition`. Access date: July 2023.

[3] `https://huggingface.co/nateraw/vit-age-classifier`. Access date: July 2023.

[4] `https://huggingface.co/oschamp/vit-artworkclassifier`. Access date: July 2023.

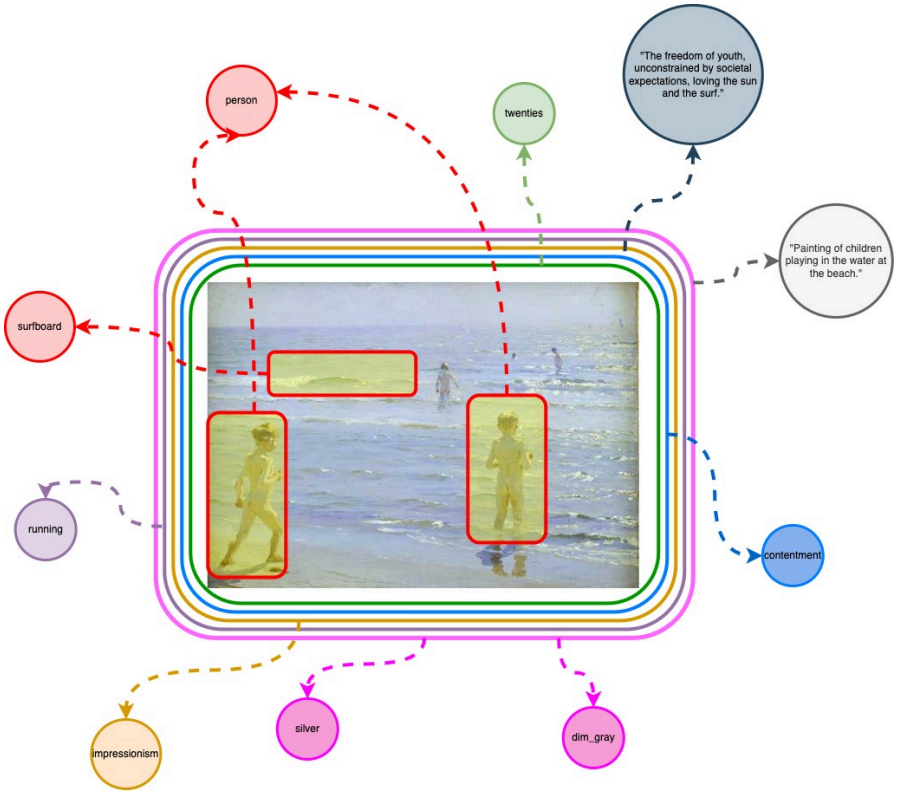


Figure III.2.3: Perceptual semantics extracted from each image in the ARTstract dataset using automatic annotators, primarily pre-trained deep learning models.

Color Detection Dominant colors were extracted from images using the ColorThief library.[5] Then, we converted the RGB values into standardized CSS3 web color names using the Webcolors Python library.[6] By measuring the Euclidean distance in a three-dimensional color space, we found the closest CSS3 web color match for each color. A closest color match was performed to find the most suitable ConceptNet concept for each color (e.g., the webcolor tag *dark goldenrod* was assigned to ConceptNet node `conceptnet:goldenrod`. If the Euclidean distance was below the threshold (set to 50), the corresponding ConceptNet concept was assigned; otherwise, it was labeled as `Unknown`. The final output included up to five RGB values, web color names, and associated ConceptNet concepts for each image.

[5] https://github.com/lokesh/color-thief. Access date: July 2023.
[6] https://webcolors.readthedocs.io/en/latest/. Access date: July 2023.

Emotion Detection For the task of detecting emotions conveyed in images, we utilized a specialized image emotion classifier pre-trained on the Artemis dataset[7]. The model classifies emotions into nine distinct categories, including *amusement*, *awe*, *contentment*, *excitement*, *anger*, *disgust*, *fear*, *sadness*, and *something else*. We matched these detected emotions with the ConceptNet concepts using the same word (e.g., the *amusement* emotion tag was assigned to the ConceptNet node `conceptnet:amusement`). To provide a comprehensive analysis of emotional content in the artwork, we retained the emotion with the highest probability for each image.

Human Presence Detection To determine the presence of humans in images, we employed a fine-tuned logistic regression classifier based on the 'ViT-B/32' variant of the CLIP (Contrastive Language-Image Pretraining) model.[8] The model was pre-trained for the fashion domain on the DeepFashion v1 dataset to classify images as either *True* for the presence of humans or *false* for the absence of humans. We matched *true* predictions with the ConceptNet node `conceptnet:human` and *False* predictions with the ConceptNet node `conceptnet:nonhuman`. For each image, we retained the presence classification with the highest probability.

Image Captioning For the task of image captioning, we employed we employed a pre-trained BLIP model.[9] This model is based on the BLIP (Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation) framework, utilizing a ViT large backbone for vision-language understanding and generation. It was pre-trained on the COCO dataset, a benchmark dataset for image captioning. The model generates descriptive captions for input images.

Object Detection For object detection, we harnessed the power of the DETR architecture using a pre-trained model.[10] We kept objects detected with a confidence threshold higher than 0.4. To enhance the semantic understanding of these detected objects, we employed ConceptNet. Each object was meticulously matched to ConceptNet concepts by its name, creating valuable connections between our object detections and the wealth of knowledge encapsulated in ConceptNet.

[7] Provided by the authors of the Artemis dataset [2].

[8] `https://huggingface.co/adhamelarabawy/fashion_human_classifier`. Access date: July 2023.

[9] `https://huggingface.co/Salesforce/blip-image-captioning-largemodel`. Access date: July 2023.

[10] `https://huggingface.co/Salesforce/facebook/detr-resnet-50`. Access date: July 2023.

III.2.3.2 Analysis of ARTstact's Perceptual Semantics

Inspired by the acquired embodiment experiments conducted in Chapter III.1, we performed co-occurrence analyses on the perceptual elements most frequently and contextually relevant to each target AC cluster. This analytical approach allowed us to identify common and significant elements, including objects, actions, colors, emotions, and more, associated with each AC.

Our approach[11] started with the collection of frequency data for images tagged with each AC was collected for all possible labels within each semantic perceptual category, such as objects, actions, emotions, art styles, and age tiers. We then assessed the frequency of labels within each category for each AC cluster, aiming to detect labels that were not only frequent (TF) but also distinctive across categories (IDF). This process allowed us to identify elements that were particularly significant for each AC. A similar approach was applied to analyze image captions. We began by extracting and preprocessing the caption data, including lemmatization, stop word removal, and word frequency calculation. Subsequently, we conducted co-occurrence analyses to identify significant word pairs that frequently appeared together in captions and were especially relevant to ACs. Additionally, for colors, we wanted to once again identify colors that had the most significance, not just the highest frequency. Therefore, we employed a two-fold strategy as before: asses their TF to gauge their prevalence within their respective ACs, and leverage the IDF to measure their distinctiveness across different categories. This combined approach enabled us to identify color labels that not only exhibited commonality within ACs but also set them apart from other categories.

III.2.3.3 PS-based AC Image Classification

Our methodological framework for AC image classification (Figure III.2.1) aligns with Pattern 6a, known as 'Intermediate Abstraction for Learning,' as defined in the modular design patterns for hybrid systems introduced by [38]. In essence, when presented with a raw ARTstract image, we employ a feature engineering approach to transform it into a vector of perceptual semantic labels. These representations serve as the foundation for training classical machine learning models, which are subsequently utilized for testing, performance assessment, and interpretability evaluation.

[11] The complete documentation and steps of our approach, including the code to execute it, is available in the project's GitHub repository, at https://github.com/delfimpandiani/ARTstract-KG/tree/main/ARTstract-KG_creation/ARTstract_kg_construction/stats. Last access date: February 2024.

Perceptual Feature Relevance Analyses

We wanted to leverage the extracted perceptual semantics (PS) to represent images as feature-engineered vectors. To do this, we first selected the number of transformation detectors to be $M = 6$, as we excluded human presence and image captions since they did not function as detected labels in the same way as the others (for instance, image captions consisted of complete sentences, and human presence was a binary true or false value). We focused on the remaining six detectors and decided how many target detections to keep track of. For example, we retained one (the most probable) action, emotion, art style, and age tier, while we kept four detected objects and four detected colors (refer to Table III.2.2 for details). While the unequal dimensionality of perceptual semantic units may introduce a bias towards colors and objects, we made this deliberate choice based on their significance in image perception and analysis. Colors and objects often serve as fundamental elements in conveying meaning, symbolism, and thematic elements within artworks. Recognizing their importance, we prioritized these features to ensure that our classification model captures essential visual cues that align with human interpretation.

PS Unit	**Function**	**Dimensionality**
Action	f_{PS_1}	$\mathbb{R}^{N_1}, N_1 = 1$
Emotion	f_{PS_2}	$\mathbb{R}^{N_2}, N_2 = 1$
Detected Objects	f_{PS_3}	$\mathbb{R}^{N_3}, N_3 = 4$
Art Style	f_{PS_4}	$\mathbb{R}^{N_4}, N_4 = 1$
Top Colors	f_{PS_5}	$\mathbb{R}^{N_5}, N_5 = 4$
Age Tier	f_{PS_6}	$\mathbb{R}^{N_6}, N_6 = 1$
	$f_{PS} = f_{PS_1} + \cdots + f_{PS_6}$	$\mathbb{R}^{N}, N = N_1 + N_2 + \cdots + N_M = 12$

Table III.2.2: Perceptual Semantic units with their respective functions and dimensionality, which are used to transform raw images into feature-engineered vectors.

Because our approach emphasizes the significance of considering perceptual elements within an image, we wanted to assess the relevance of these features when predicting abstract concepts. To do so, we employ information theory principles. Specifically, we calculate the entropy of a perceptual feature (Y) under the condition of the AC (X) evoked by the image using the following equation:

$$\mathrm{H}(Y|X) = -\sum_{x\in\mathscr{X},y\in\mathscr{Y}} p(x,y)\log\frac{p(x,y)}{p(x)} \tag{III.2.10}$$

Conditional cross-entropy, denoted as $\mathrm{H}(Y|X)$, measures how much additional information each perceptual feature contributes when predicting the abstract concept cluster X. In essence, it gauges whether knowing certain features makes the abstract concept more predictable. A higher $\mathrm{H}(Y|X)$ indicates that specific features add more information, making the concept more predictable when those features are known. Conversely, a lower value suggests that these features provide less additional predictive information about the abstract concept cluster, making it less predictable based on them. In addition to the general overview of feature importance, we calculate the conditioned cross-entropy by AC cluster, delving deeper into the concept-wise relevance of each feature.

Feature Engineered Image Representation

With this information, we created a new Perceptual Semantic (PS) image representation denoted as I_{PS}. This representation is obtained using the f_{PS} function, which transforms the raw image (I_{RAW}) into a vector space with a dimensionality of $\mathbb{R}^N$:

$$f_{PS}(I_{\mathrm{RAW}}) = I_{\mathrm{PS}} \subseteq \mathbb{R}^N, N = 12 \tag{III.2.11}$$

Initially, we preserved all original perceptual features, representing each image as a 12-dimensional vector, with each dimension corresponding to a unique feature, as illustrated in Figure III.2.10. This representation effectively translated perceptual semantics into a tabular format, where each image constituted a row, and each column represented each of the possible labels for each of the 12 features. However, when subjecting representation to classification tests (as discussed in the following section), we encountered significant challenges, leading to suboptimal performance. In response, we leveraged the feature relevance results (See Section III.2.4) to gain insights into the significance of individual features. Subsequently, we made informed decisions to retain the most relevant perceptual semantics categories, discarding the last two objects, the final color, and the age feature. This refinement led to a more streamlined 8-dimensional vector, denoted as $I_{\mathrm{PS}} \subseteq \mathbb{R}^N$, with N reduced to 8 dimensions.

Classical ML Methods Training and Testing

This newly obtained representation I_{PS} was integral to our problem formulation. After transforming it into a tabular representation where each image constituted

a row, and each column represented each of the possible labels for each of the 8 features, this representation enabled the training of classical machine learning models for making inferences based on the following probability estimation:

$$\hat{y} = \arg\max(p(y_i|I_{\text{PS}}, \theta)) \tag{III.2.12}$$

We harnessed this tabular representation to train a variety of classical machine learning models, including Bernoulli Naive Bayes, Decision Tree Classifier, Random Forest Classifier, HistGradientBoostingClassifier, and Linear Support Vector Classification through the Scikit-learn Python library [279] and implemented a Bayesian network with a manually defined structure to encapsulate the dependence relationships.

Evaluation Metrics

We applied the same set of performance metrics, which includes accuracy, precision, recall, and macro F1 score, as in the experiments detailed in Chapter II.2. We also retained the exact training and testing data splits used in the deep learning experiments presented in Chapter II.2.

Explainability Approach

To enhance the interpretability of our best-performing model's classifications, we adopted an approach in line with modular design pattern 5, "Explainable Learning System through Rational Reconstruction", again following the taxonomy by [38]. This approach leverages the detected label and the model responsible for the detection to derive instance-level explanations. Our specific focus was on the best-performing model, the Naive Bayes classifier. Given our I_{PS} image representation, the model becomes fully explainable: we can identify the features, denoted as PS_{f_n} contributing most significantly to the highest probability of a particular class:

$$p(y_i|I_{\text{PS}}, \theta) = p(y_i|I_{\text{PS}_{f_0}}, \theta) + p(y_i|I_{\text{PS}_{f_1}}, \theta) + \cdots + p(y_i|I_{\text{PS}_{f_N}}, \theta) \tag{III.2.13}$$

$$p(y_i|I_{\text{PS}}, \theta) = \sum_{n=0}^{N} p(y_i|I_{\text{PS}_{f_n}}, \theta) \tag{III.2.14}$$

III.2.4 Results

III.2.4.1 ARTstract's Perceptual Semantics

The co-occurrence and relevance analyses and visualizations of ARTstract's Perceptual Semantics provide insights into elements that are not only frequent but also especially prominent within the images tagged with a specific abstract concept (AC). The results are visualized through wordclouds and color palettes created from the most relevant web colors for each AC (examples are shown in Figures III.2.6 and III.2.9, with additional examples available on our GitHub repository). The findings reveal intuitive and meaningful patterns. For instance, the most relevant actions for the AC *comfort* include 'eating' and 'hugging,' while 'running' and 'sitting' are prominent for *freedom*. The analysis of relevant objects in the ARTstract dataset also exposes interesting biases. Images tagged as *comfort* are associated with objects such as 'potted plants,' 'vases,' and 'couches,' suggesting a prominence of images related to nature of home areas tagged with *comfort*, whereas objects like 'bird' and 'kite' are more relevant to *freedom*, suggesting the presence of a bias towards these objects in images tagged with it. Moreover, the examination of relevant colors, achieved by mapping pixels to web color names and identifying the most relevant ones through TF-IDF analysis, offers intriguing results. For example, the relevant colors for *freedom* closely resemble the United States flag, aligning with the presence of the word 'america' within the cluster definition. Overall, these findings serve as a valuable resource for diverse applications across domains, suggesting that the extracted perceptual semantics may contain sufficient information to classify images based solely on these visual features.

III.2.4.2 Perceptual Feature Relevance Analyses

The results of this cross-entropy analysis (Figure III.2.10) reveal that all perceptual features carry information when conditioned on the cluster, but their importance varies. Colors and the presence of first objects stand out as highly informative, whereas the fourth object, emotion, and age features contribute less to the predictability of abstract concept clusters. This provides a comprehensive overview of the relative importance of different feature categories in our analysis, allowing for a bird's eye view of their contributions as a whole.

In addition to the general overview of feature importance, we calculate the conditioned cross-entropy by AC cluster (Figure III.2.11), delving deeper into the concept-wise relevance of each feature. We find that emotion features consistently play a significant role across most AC clusters, with a notable exception in the *comfort* cluster, likely due to the availability of a larger dataset, resulting in higher variance. *Safety* and *freedom* clusters are observed to be less reliant on art style,

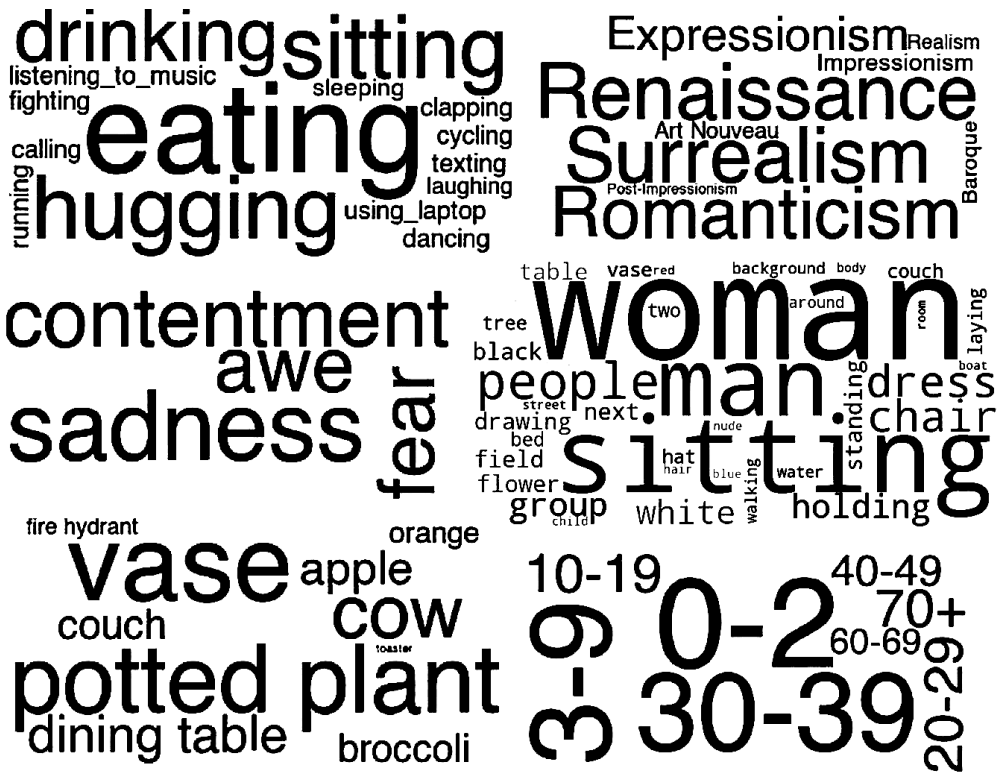


Figure III.2.4: *Comfort* wordclouds.

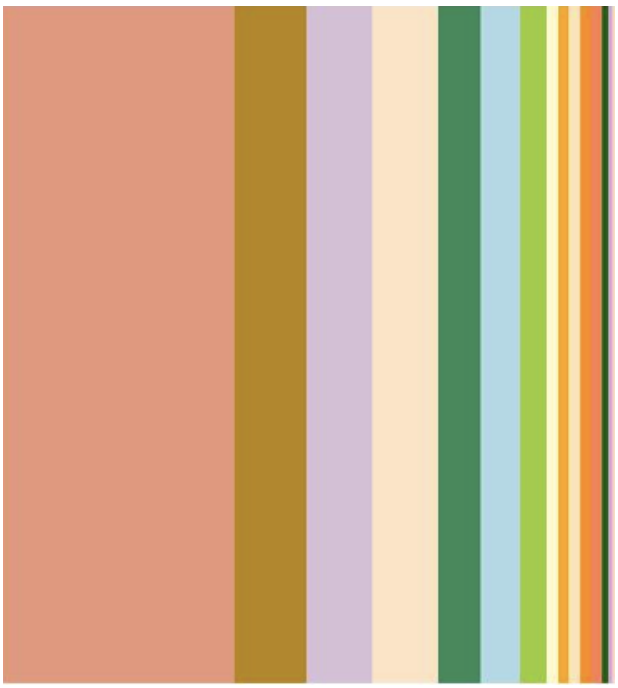


Figure III.2.5: *Comfort* relevant colors.

Figure III.2.6: Perceptual semantics most relevant to the AC *comfort*, based on co-occurrence data obtained from the complete ARTstract dataset. Top, clockwise from top left: wordclouds for actions, art style, caption words, age tiers, detected objects, and emotions. Bottom: most relevant colors for *comfort*.

Figure III.2.7: *Freedom* wordclouds.

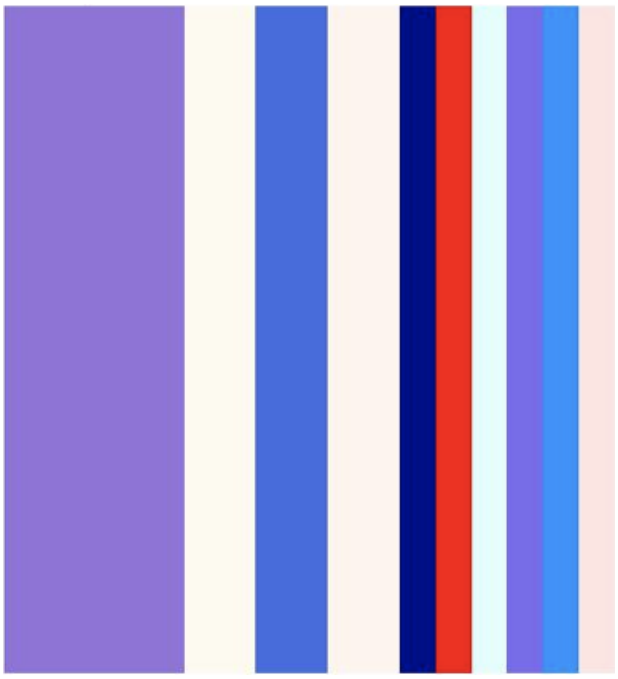


Figure III.2.8: *Freedom* relevant colors.

Figure III.2.9: Perceptual semantics most relevant to the AC *freedom*, based on co-occurrence data obtained from the complete ARTstract dataset. Top, clockwise from top left: wordclouds for actions, art style, caption words, age tiers, detected objects, and emotions. Bottom: most relevant colors for *freedom*.

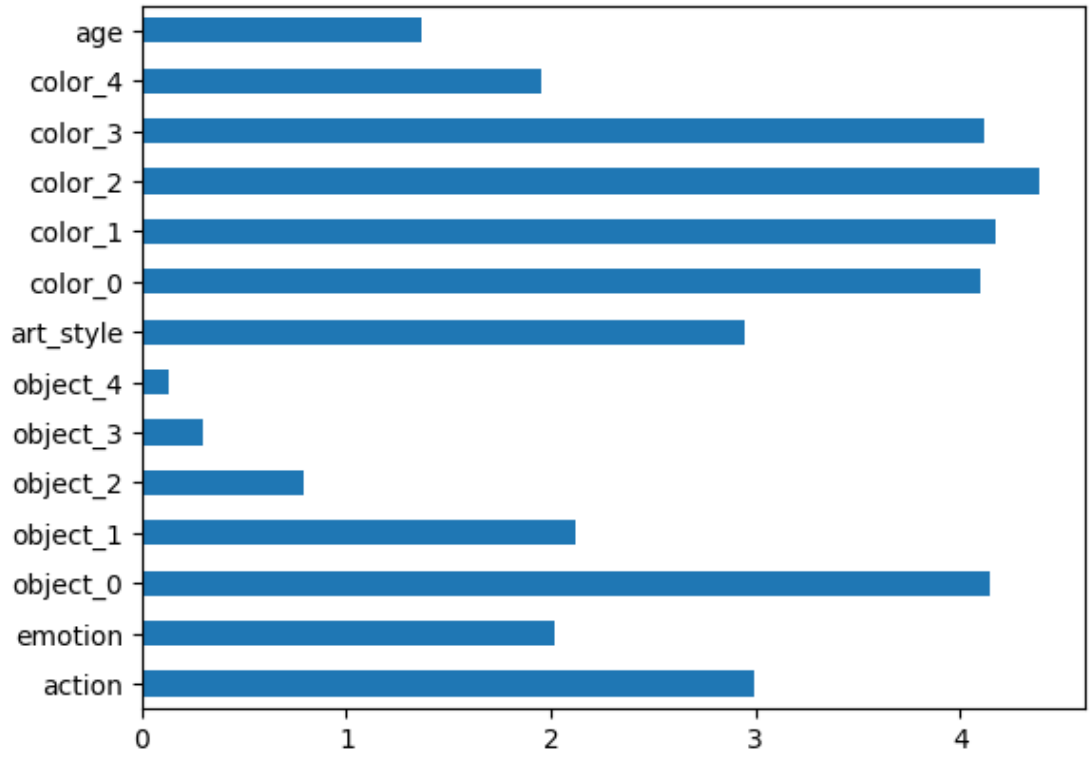


Figure III.2.10: Conditioned Cross-Entropy

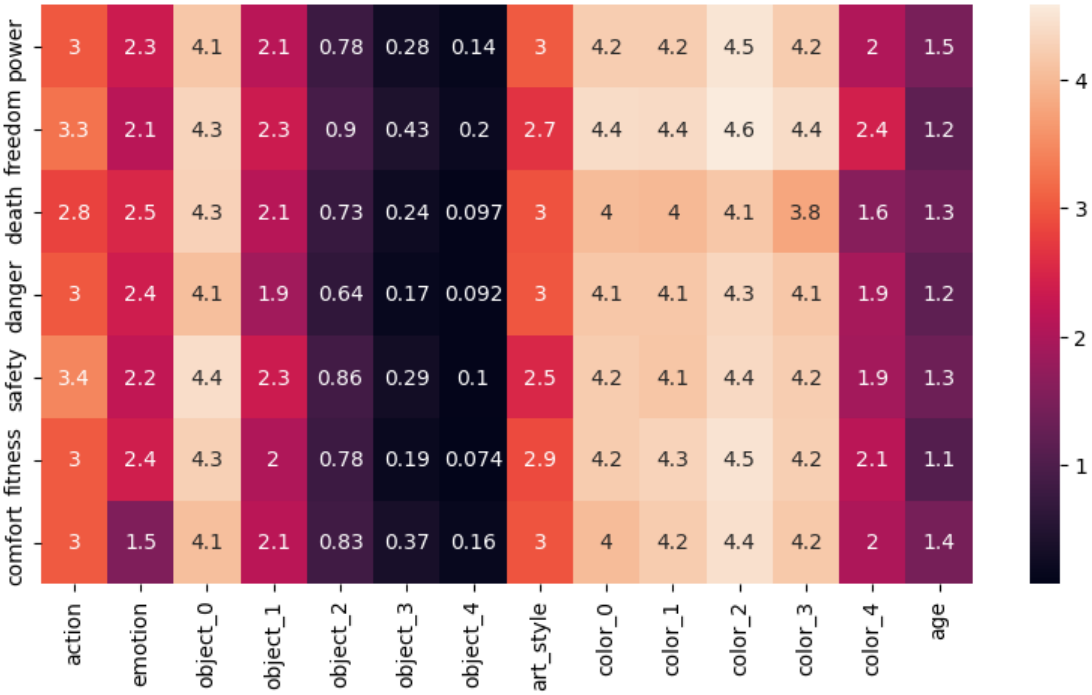


Figure III.2.11: Conditioned Cross-Entropy by Cluster

indicating a reduced contribution from art style features for these specific clusters. Age features exhibit uniform lower relevance across all clusters on average. We note that objects 2, 3, or 4 show a similarly low relevance, which can be attributed to their absence in some samples. This suggests that, while they may seem unimportant, their presence in certain cases could indeed be significant, emphasizing the conditional nature of their relevance

Overall, this analysis provides quantitative evidence, consistent with the perceptual semantics paradigm, revealing the contribution of perceptual features to Abstract Concept (AC) predictability. It demonstrates that various feature categories play varying roles in AC prediction, emphasizing the conditional nature of these dependencies across AC clusters.

III.2.4.3 Performance

Table III.2.3 provides a comprehensive comparison of the performance metrics for various classical machine learning methods applied to the AC image classification task. Each row corresponds to a specific method, and the columns show the respective accuracy, macro F1, and weighted F1 scores, along with the support for each method. Notably, the macro F1 score is particularly relevant as it accounts for the harmonic mean of precision and recall across all AC clusters.

The results highlight that, among the studied methods, Naive Bayes stands out with a macro F1 score of **0.24**, demonstrating its ability to balance precision and recall effectively across different AC clusters. It achieved an accuracy of 0.44 and a weighted F1 score of 0.40, showcasing its solid overall performance. These results are further visualized in Figure III.2.12, which illustrates the macro F1 scores for each abstract concept cluster using the different classical ML methods. Detailed metrics for each class according to each of the classical ML methods can be found in the Appendix, in Section V.1.5.3.

Method	Scores			Support
	Accuracy	Macro F1	Weighted F1	
Decision Tree	0.35	0.20	0.34	1492
Random Forest	0.44	0.20	0.38	1492
XGB	0.45	0.20	0.39	1492
SVM	0.45	0.20	0.38	1492
Bayesian Network	0.42	0.20	0.37	1492
Naive Bayes	0.44	**0.24**	**0.40**	1492

Table III.2.3: Comparison of performance metrics for various classical machine learning methods used in abstract concept image classification, including accuracy, macro F1 score, and weighted F1 score.

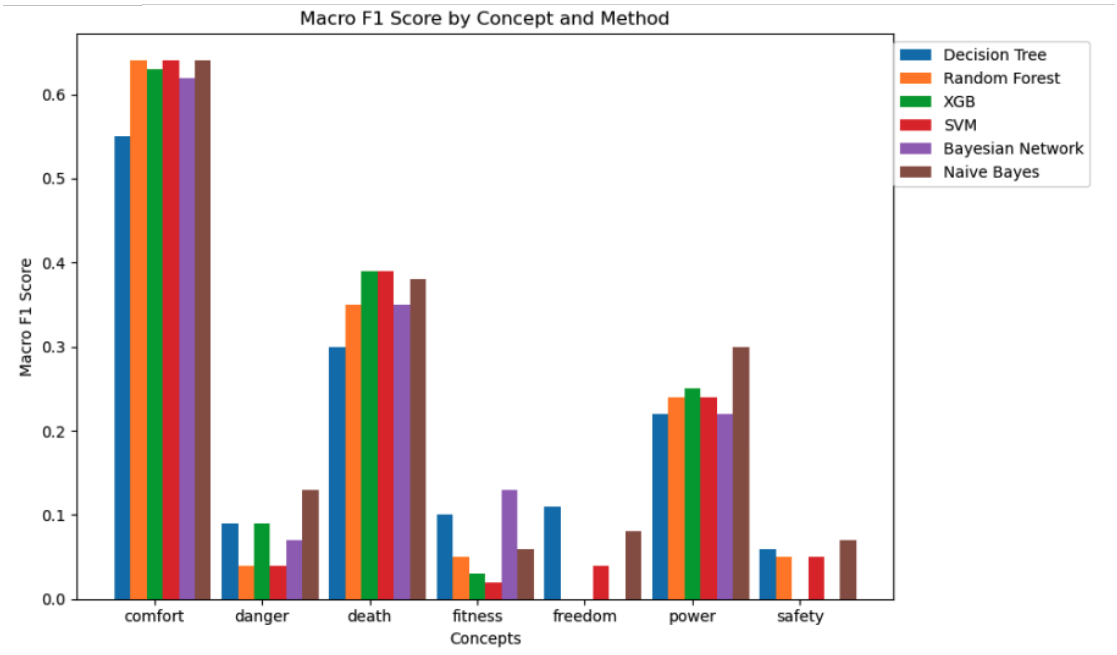


Figure III.2.12: Macro F1 scores for each of the classical ML methods on each of the AC clusters.

Comparison with DL performance

Table III.2.4 compares the performances of top three best-performing classical machine learning (ML) methods and of the top three deep learning (DL) models. The DL models, VGG-16, ResNet-50, and ViT exhibit better levels of accuracy, and ViT outperforms all methods in all metrics. However, when we shift our focus to the F1 score, a measure that gauges both precision and recall, the Naive Bayes method not only performed on par with the two CNNs but in one case outperformed it.

Method	Scores			Support
	Accuracy	Macro F1	ML/DL	
XGB	0.45	0.20	ML	1492
SVM	0.45	0.20	ML	1492
Naive Bayes	0.44	**0.24**	ML	1492
VGG-16	0.47	0.23	DL	1492
ResNet-50	0.48	**0.24**	DL	1492
ViT	0.51	***0.30***	DL	1492

Table III.2.4: Comparative analysis of the best performing classical ML and DL models. The top-performing model, measured by the F1 score (ViT), is highlighted in both bold and italics. The second-best performing models (ResNet-50 and Naive Bayes) are denoted in bold.

III.2.4.4 Explainability

Incorporating an explainability approach into our workflow, we focused on our best-performing model, the Naive Bayes classifier, to provide transparent insights into its classifications. By leveraging our I_{PS} image representation, we identified the perceptual semantic features PS_{f_n} that most significantly contributed to the model's high probability for a specific class. This instance-level explanation approach facilitated a deeper understanding of the model's reasoning. To exemplify, we present a specific case in Figure III.2.14, where we present an exemplary test image alongside the output of our explainability approach.

The ground truth label assigned to the image is *danger*. However, the model classifies this image as *death* instead, exhibiting high confidence (Probability: 0.7899). This instance sheds light on the complexity of the classification task, suggesting that it could be interpreted not merely as a single-label multiclass problem but rather as a multi-label multi-class one. This implies that the evaluation

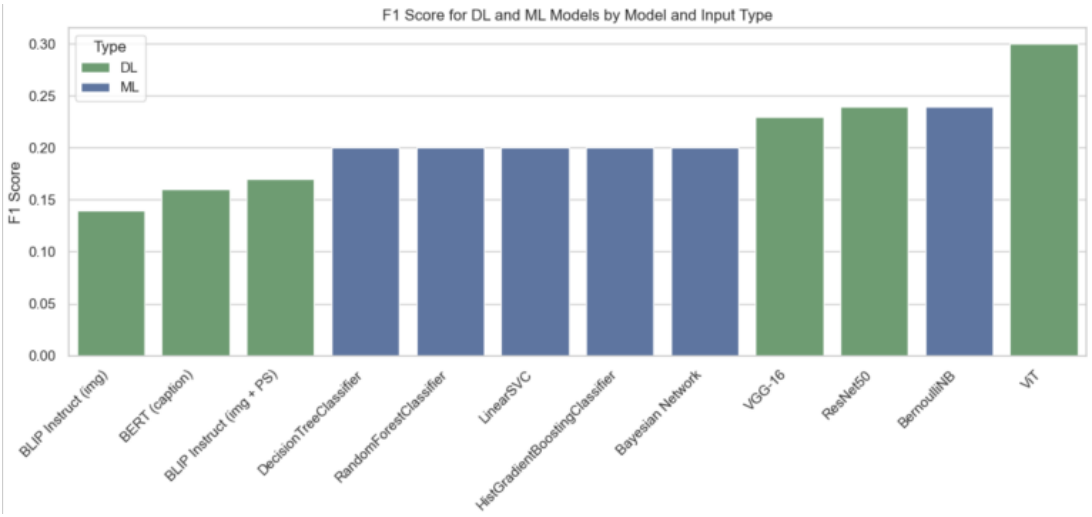


Figure III.2.13: Macro F1 scores for each of the classical ML methods as well as the deep models. Naive-Bayes shows competitive performance.

and performance metrics should be extended to encompass lower-ranked concepts as well. The accompanying explanation provides a breakdown of the perceptual features contributing to this classification, elucidating the model's reasoning. The most relevant features include object detection for "person" (0.324), the action "sleeping" (0.186), the emotion "fear" (0.091), and specific color detections like "darkgray" (0.062) and "lightslategray" (0.050), among others.

III.2.5 Discussion

In this chapter, we have delved into the crucial role of perceptual features in image classification based on ACs. We introduce a novel paradigm known as *perceptual semantics*, which involves transforming raw image data into an understandable vector representation. This representation captures tangible perceptual units, effectively bridging the gap between raw pixels and ACs in a more interpretable manner compared to feature vectors extracted from deep learning models. Our approach automatically extracts these semantic units, including action, age tier, art style, top colors, emotion, human presence, image caption, and detected objects. Each unit is linked to a ConceptNet node, enriching the semantics associated with images. Additionally, we conducted co-occurrence analyses of perceptual elements within each target AC cluster, offering valuable insights into the most relevant perceptual features for each AC. To gauge the importance of these fea-

Class Rank 1: *death* (Probability: 0.7899)

PS Unit	Label	Probability
object_0	**person**	0.3245
action	**sleeping**	0.1858
emotion	**fear**	0.0914
color_2	**darkgray**	0.0619
color_3	**lightslategray**	0.0501
art_style	**Renaissance**	0.0354
object_1	**horse**	0.0295
color_0	**tan**	0.0088
color_1	**darkslategray**	0.0088

Figure III.2.14: Instance-level explanation for a test image demonstrates the power of our explainability method, providing a detailed breakdown of the perceptual features influencing the model's classification decisions. This example highlights how our explainability approach unveils the inner workings of the model, promoting transparency and providing valuable insights into the model's reasoning for AC classification.

tures, we employed information theory principles, using conditioned cross-entropy calculations at both a general and concept-specific level within AC clusters, and gained a holistic and contextual understanding of the significance of perceptual features.

We then proposed a method for image classification using the extracted perceptual semantics (PS). This approach involves selecting six transformation detectors and leveraging the extracted information to create a new Perceptual Semantic (PS) image representation. Our methodology aligns with modular design pattern 6a from the taxonomy of hybrid systems by [38], emphasizing intermediate abstraction for learning. This method builds upon the insights obtained from previous sections of the dissertation, making the problem more interpretable and facilitating the identification of specific perceptual features contributing to class predictions.

III.2.5.1 Perceptual Feature Relevance

The cross-entropy analysis depicted in Figure III.2.10 shows that all perceptual features carry information when conditioned on the cluster, but their significance differs. Notably, colors and the presence of initial objects emerge as highly informative, while the fourth object, emotion, and age features contribute less to the predictability of abstract concept clusters. This offers a comprehensive view of the relative importance of distinct feature categories in our analysis. Additionally, the conditioned cross-entropy by AC cluster (Figure III.2.11) delves into concept-

specific feature relevance, offering quantitative proof of how perceptual features contribute to Abstract Concept (AC) predictability, emphasizing the conditional nature of these dependencies across AC clusters.

III.2.5.2 Performance

The results highlight the impressive performance of the Naive Bayes classifier with a macro F1 score of 0.24. What makes this accomplishment particularly noteworthy is that classical ML models, including Naive Bayes, were exclusively supplied with only the 8 perceptual semantic labels for each image, without any additional information such as deep features or pixel-level data. This underscores the significant potential of interpretable ML techniques, as they exhibit a well-balanced balance between precision and recall, even when working with minimal input data. Notably, our comparative analysis in Table III.2.4 indicates that Naive Bayes performs on par with or even outperforms deep learning models like VGG-16 and ResNet-50. This finding challenges the conventional notion that deep learning models are the solution for all image classification tasks, underlining the practicality and effectiveness of interpretable machine learning methods in this context. Our explainability approach, as exemplified by Figure III.2.14, provides transparent insights into the model's decision-making process, enhancing the overall interpretability and user trust in model outputs.

III.2.5.3 Interpretability

In our approach, we introduced explainability into our workflow, focusing on the Naive Bayes classifier, our best-performing model. Leveraging our fully interpretable image representation, we were able to pinpoint the most influential perceptual semantic features contributing to the model's classifications. We exemplify this in a specific case, where the model confidently categorizes an image as *death* instead of its actual class, *danger*. The accompanying explanation dissects the key features behind this decision, including object detection, actions, emotions, and colors. The example illustrates how our explainability approach dissects model decisions and highlights the influence of particular perceptual features in the classification process. The identification of these relevant features provides valuable insights into the model's thought process when determining abstract concept labels for test images. The significance of this approach becomes evident when comparing it to deep learning models. While deep learning models might offer similar classification performance, their decision-making process often remains a 'black box.' In contrast, our approach unveils the concrete factors leading to specific predictions, thus enhancing transparency and enabling users to make more informed and interpretable decisions based on model outputs.

Some of the key findings and lessons that emerge from this work include:

Perceptual Semantics (PS) Automatic Extraction A key contribution of this chapter is the delineation and automation of Perceptual Semantics (PS) extraction from images, replicating human vision and high-level visual sensemaking. This achievement stands out by eschewing human-labeled concrete data and embracing a fully automated, consistently applied process to all images. This approach not only bolsters scalability but also underscores the transformative potential of automation in image analysis. By exclusively relying on state-of-the-art detectors for extraction, our research streamlines the process, providing insights into how contemporary machine vision systems perceive visual content. The broad spectrum of perceptual semantics considered, spanning from emotions to art styles, adds depth and nuance to image analysis. This work advances computer vision by mirroring contemporary machine vision systems, opening vast possibilities for applications in art, culture, advertising, and marketing.

Innovative Image Representation This work introduces a novel approach to image representation, emphasizing the transformation of raw image data into interpretable vector representations. This method opens up new possibilities in image analysis and classification, which can allow for a more cognitive-aligned understanding of image content beyond traditional methods. This approach has significant potential applications across various domains, from art and culture to cognitive science and machine learning.

Interpretability vs. Complexity The choice of using classical machine learning, particularly Naive Bayes, raises important questions about interpretability and complexity in image classification. By opting for an interpretable model over deep learning approaches, the chapter highlights the trade-offs and implications of this decision. It underscores the importance of interpretable AI models and their role in improving trust and user adoption. Critically, our findings challenge the conventional notion that deep learning models are the sole solution for image classification tasks, underlining the practicality and effectiveness of interpretable machine learning methods in this context. Our explainability approach, as exemplified by Figure III.2.14, provides transparent insights into the model's decision-making process, enhancing the overall interpretability and user trust in model outputs. This observation calls for a reevaluation of the conventional belief in complex deep learning models as the default choice, emphasizing the practicality and effectiveness of interpretable machine learning methods in the field of abstract concept image classification.

Interdisciplinary Research This section represents a unique amalgamation of computer vision, cognitive science, and cultural image analysis, yielding novel insights that reshape our approach to achieving interpretability in computer vision. The interdisciplinary nature of this research has significant implications for knowledge integration, emphasizing the importance of collaborative approaches in advancing the field. This work demonstrates that when various disciplines converge, it is possible to uncover new dimensions of understanding, ultimately leading to more robust and effective solutions for complex problems in computer vision and image analysis.

III.2.5.4 Future Directions

While our research has uncovered essential lessons and avenues for exploration, it also highlights promising directions for future research, including:

Situating the Perceptual Semantics (PS) The automation of extracting perceptual semantics from images marks a significant achievement, offering advantages such as efficiency and scalability. However, it also presents challenges, including potential biases in automated detection. To further our research, we should delve into contextualizing the perceptual semantics using information about the strength of annotations, the artificial annotator used, the model architecture, dataset pretraining details, and other relevant contextual information. This avenue of research is particularly promising for bridging the cultural gap, especially in light of the insights gained from Chapter IV.2.

Further Semantic Enrichment and Exploitation of Commonsense Knowledge The assignment of perceptual semantic labels to ConceptNet nodes significantly enriches the semantics associated with images, resulting in a deeper understanding of image content. However, our methods have yet to fully exploit this semantic enrichment. The existence of this valuable semantic content serves as a reminder that the extracted semantics can and should be further reasoned over, drawing lessons from Chapter IV.1 to guide our exploration.

Comparability Across Datasets and Potential for Generalization Our decision to apply a consistent process to all images within the ARTstract dataset underscores the generalizable nature of the approach. We should explore the potential for applying this approach to other image datasets and domains beyond the context of art images or perhaps beyond the scope for AC image classification, expanding the scope of our research.

Exploring Visual Features for Image Classification The information theory results highlight the significance of colors in image classification. This suggests that incorporating specific color data directly from image pixels, in addition to the PS, could lead to improved performance for classifiers. Consideration should be given to experiments that combine PS with deep features from models like Vit and VGG, which excel at capturing raw perceptual features such as colors, shapes, and lines.

Co-occurrence Analysis and Abstract Concept Descriptions The results from analyzing ARTstract's perceptual semantics (e.g., Figures III.2.6 and III.2.9) suggest that the data we have collected can be statistically analyzed for the frequencies and relevance of specific features concerning specific AC clusters, enabling the identification of common-sense descriptions of abstract concepts. Further research is warranted on how to formally represent ACs based on these particularly relevant features. These multimodal AC representations, combining textual and color information, could find applications in various domains beyond AC image classification, but also for image generation and psychological or cognitive research.

Balance Representation of Perceptual Features The unequal dimensionality of perceptual semantic units likely introduces a bias towards colors and objects. Future iterations of our methodology may explore strategies to address this bias more effectively, aiming for a more balanced representation of perceptual features in image classification tasks.

Explainability User Assessment While the findings on explainability are promising, integrating evaluations from experts in art history could provide invaluable insights. Future research should prioritize conducting user studies, particularly involving art history experts, to assess the clarity, comprehensibility, and usefulness of the explainability method.

III.2.6 Conclusions

Through the delineation and automation of the *perceptual semantics* (PS) extraction process, we open the door to replicating the cognitive aspects of human high-level visual sensemaking. This approach not only ensures scalability but also broadens our understanding of how machines interpret visual content, allowing for statistical analyses to illuminate the relationship between ACs and perceptual elements. Critically, our findings challenge the conventional notion that deep learning (DL) models are the sole solution for image classification tasks, underlining the practicality and effectiveness of keeping the DL paradigm close to the concrete

levels and using more interpretable machine learning methods to build on top. Our explainability approach, as exemplified by Figure III.2.14, provides transparent insights into the model's decision-making process, enhancing the overall interpretability and user trust in model outputs. This paradigm reveals a nuanced landscape where feature engineering and traditional machine learning approaches can provide explainable decisions for higher-level tasks without significantly sacrificing accuracy. Overall, this interdisciplinary work bridges computer vision, cognitive science, and cultural image analysis, unveiling new horizons in the pursuit of explainable AI. As we explore further directions, we recognize the need to contextualize perceptual semantics, exploit semantic enrichment, and extend our approach to various datasets and tasks. In doing so, we pave the way for the integration of engineered perceptual semantics in image analysis, enhancing the cognitive alignment between machines and humans, with far-reaching implications for fields ranging from art and culture to AI and beyond.

Part IV

Reifying and Reasoning with Knowledge Graphs

Chapter IV.1

Interpretable Bridging of Visual Data and Linguistic Frames

Summary Chapters II.2 and III.1 offer compelling evidence regarding the role of concrete elements, particularly depicted objects, in connecting raw pixel data with ACs, emphasizing the potential of exploiting concrete semantics. Considering insights from cognitive science on the importance of distributional linguistic and commonsense knowledge in AC representation (Chapter I.3), a further way to exploit them emerges: using linguistic and common-sense reasoning to ascend from concrete visual elements to abstract knowledge by identifying frames, pivotal cognitive structures. Despite the significance of frame evocation for visual and multimodal sense-making, a notable lack of data-driven tools to automate this process exists. This chapter addresses the challenge by automating the reasoning process on the concrete semantics of visual data to unveil associations with high-level linguistic frames. Through ontology-based knowledge engineering techniques, we provide an explainable framework for identifying "framal visual manifestations." The chapter conducts a deep ontological analysis of the Visual Genome image dataset [233] and introduces the Visual Sense Ontology (VSO). To enrich the dataset, we present a fully interpretable framal knowledge expansion pipeline that extracts and links linguistic frames, including values and emotions, to images, using multiple linguistic resources for disambiguation. Moreover, we introduce the Visual Sense Knowledge Graph (VSKG), enhancing the accessibility and comprehensibility of Visual Genome's multimodal data through SPARQL queries. VSKG encompasses data on frame visual evocation, enabling advanced explicit reasoning, analysis, and sensemaking. Our work advances the automation of frame evocation and multimodal sense-making while maintaining interpretability and transparency.

IV.1.1 Introduction and Background

In communication studies, sensemaking is a process of categorization and labeling to bring stability to an individual's experiences, requiring the selection and transformation of certain elements from everyday experiences into abstract social categories when faced with something unfamiliar [170]. It has been argued that verbal discourse alone is not an accurate representation of communication in contemporary society [39, 40] and, instead, human communication involves multiple modes, including visual elements, working together to create meaning. As such, multimodal sensemaking is a crucial aspect of human cognition, enabling us to synthesize and give meaning to our experiences. This process involves integrating knowledge from different modes, including visual, linguistic, physiological, and auditory, to perceive, reason, learn, and take action. As a result, automated multimodal reasoning has emerged as a promising area of research in various fields, including Cultural Heritage and Human-Computer Interaction.

In recent years, automatic multimodal analysis has gained popularity, but with an almost exclusive focus on the use of DL techniques. [156] provides a comprehensive survey on deep multimodal representation learning, which aims to narrow the heterogeneity gap among different modalities in the utilization of ubiquitous multimodal data. Additionally, DL models such as CNNs have been used in CV for a range of related tasks, including image classification, object detection, and language modeling [212, 294, 347]. Within the field of CV, situational analysis aims to automatically detect commonsense semantics, including actions, activities, roles, and interactions between objects, in visual situations, scenes, and events. Within situational analysis, explicit tasks include abstract reasoning, which deals with global semantic tasks based on logic, such as work by [337], and Grounded Situation Recognition (GSR) [280], which involves producing structured semantic summaries of images, including identifying the primary activity, entities involved in the activity with their roles, and bounding-box groundings of entities. More related work includes Human Activity Recognition (HAR) systems [36] and multi-feature, multi-modal, and multi-source event recognition [5], which includes work on recognizing specific social and cultural events.

The aforementioned DL models are often used for the task of linking lexical knowledge to visual data. However, the problem of adopting black-box approaches to sensemaking tasks is twofold: (i) their black-box nature limits the understanding of their decisions and inner workings, such that it is not possible to completely backtrack the determinant parameters, for example, of a classification task; and (ii) a considerable portion of semantics e.g. the meaningful co-occurrence of several elements co-participating in the same event, is flattened to a single parameter/label. A feasible alternative approach to DL architectures encompasses the use of ontologies, knowledge base systems, and symbolic reasoning. Ontologies, as in-

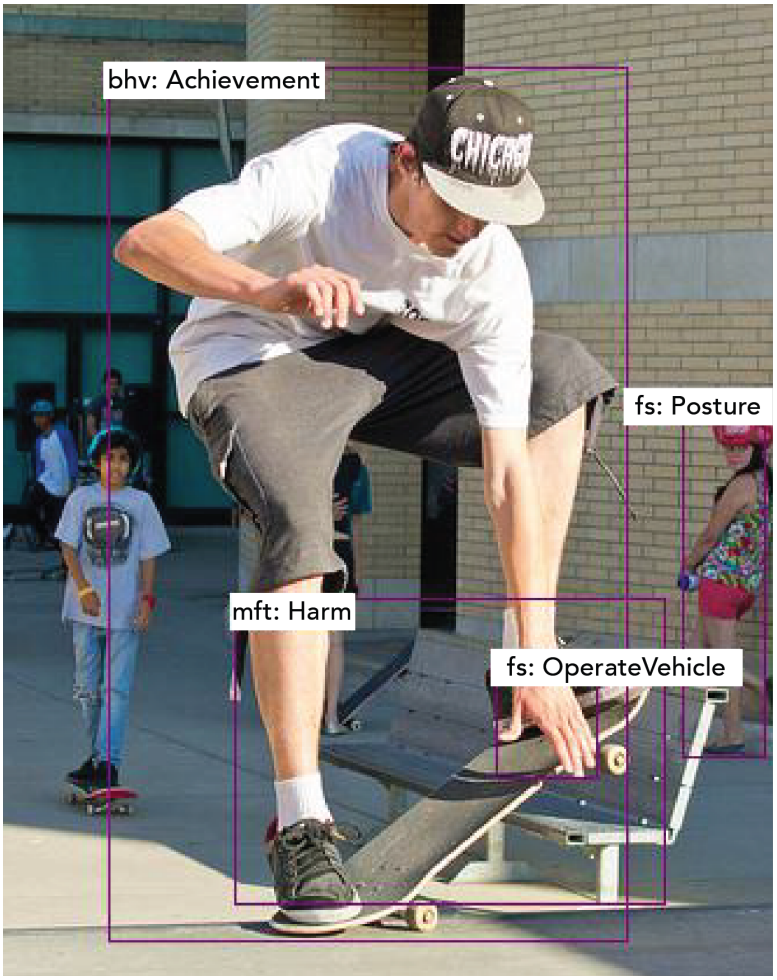


Figure IV.1.1: Framal visual instantiations automatically extracted with our pipeline from a Visual Genome image. The pipeline detects linguistic frames from the image's region descriptions and connects them to their visual manifestations. Original image from the Visual Genome dataset [208].

tended in computer science, are formal and explicit representations of knowledge in a certain domain [153, 151]. The use of ontology-based knowledge and data can enhance both the performance and the explainability of automatic reasoning over data for decision-making tasks, as well as for knowledge retrieval.

Symbolic knowledge previously used to aid the performance and explainability of visual learning tasks include the use of logic rule explanation methods [3], as well as the incorporation of external knowledge via ontology-based KGs, leading to the proliferation of multi-modal KGs (MMKGs) (survey by [407]). These include BabelNet KG, which integrates many sources and covers a wide range of

languages [263], BabelPic [65] a hand-labeled image dataset explicitly targeting non-concrete concepts, and the CommonSense Knowledge Graph (CSKG) proposed by [176]. CSKG consolidates commonsense knowledge from seven different sources and provides useful embeddings for downstream reasoning and pre-training of language models. However, most of these ontologies are not described in depth and are also not publicly available for SPARQL querying. As such, there is a scarcity of frameworks that allow the direct querying of and reasoning over image data, limiting their potential for multimodal knowledge enrichment.

We argue that the kinds of situational knowledge targeted by deep architectures can be detected in an automatic, explainable, and more human-like fashion via the explicit integration of visual data and linguistic frames. This is because reasoning via linguistic descriptors, with features of encyclopedic knowledge representing cognitive phenomena, has been successfully performed by Frame Semantics methodologies [359]. A frame is, in Fillmore's broad definition [120], a cognitive representation of typical features of a situation. Frames are structures that formalize the network of meaning in semantic roles participating in a certain situation. This network of semantic references to triggers of meaning is usually called "activation" or "evocation" of some frame. As such, a different and potentially more explainable approach to automatic multimodal sensemaking is the automatic detection of frame evocation from visual data, which involves the integration of various sources of knowledge, including lexical and commonsense resources. However, there is a scarcity of data-driven approaches and tools that integrate different modes of frame-based knowledge for automatic multimodal sensemaking.

To address this scarcity, this work proposes a multimodal integration and reasoning pipeline that provides resources, algorithms, tools, and methods for automated multimodal inferences. Specifically, we focus on the explicit and formal integration of two important resources, Visual Genome (VG) [208], a large annotated image dataset, and Framester [131], a linked open data graph resource that provides linguistic and factual knowledge.

The objective of this work was to develop a full pipeline that allows, for a VG image of choice, the automatic modeling, implementation, and publication of a semantic web KG (in RDF) containing multimodal data, including visual frame evocations (see Fig. IV.1.1). Our approach prioritizes ontology-based knowledge engineering, as ontologies offer a structured means of representing knowledge and the connections between concepts. By utilizing ontologies, we can seamlessly integrate data from diverse sources and facilitate reasoning about the information. As such, this work introduces the Visual Sense Ontology (VSO) and the Visual Sense Knowledge Graph (VSKG), a linked data KG that contains multimodal (factual, linguistic, and visual) knowledge.

Our work provides a valuable contribution to the field of multimodal sensemak-

ing, by presenting a data-driven approach to automatically performing frame-based inferences based on multimodal data. By providing a comprehensive framework for integrating and reasoning with multimodal knowledge, the resulting linked data KG has the potential to support a wide range of applications, from enhancing human-computer interaction to advancing the state of the art in knowledge representation, CV, and natural language understanding. The associated GitHub repository[1] and project website make our work accessible to researchers and practitioners in this field.

IV.1.2 Resources and Tools

In this section, we briefly introduce the main resources and tools reused, in particular the Visual Genome dataset and the Framester ontological hub. The FRED tool [132] to generate KGs from natural language is also described.

IV.1.2.1 Visual Genome

Visual Genome (VG) [208] is an annotated image dataset containing over 108K images where each image is annotated with an average of 35 objects, 26 attributes, and 21 pairwise relationships between objects. Regarding relationships and attributes as first-class citizens of the annotation space, in addition to the traditional focus on objects, VG's annotations represent the densest and largest dataset of image descriptions, objects, attributes, relationships, and question-answer pairs. The Visual Genome dataset is among the first to provide detailed labeling of object interactions and attributes, providing a first step of grounding visual concepts to language by canonicalizing the objects, attributes, relationships, noun phrases in region descriptions, and question-answer pairs to WordNet synsets.

IV.1.2.2 Framester

The notion of "frame" refers to Fillmore's Frame Semantics [119, 120]: frames are schematic formalizations, in the form of N-ary relations, of commonsense knowledge cognitive representations about entities and situations. The grounding assumption is that the semantics of a certain entity cannot be understood without considering a minimum context in which some meaning is situated. FrameNet is the resource originally formalizing this knowledge, structuring each frame as composed by some necessary frame elements, namely, semantic roles, and linking lexical units (LU), namely, each piece of linguistic material commonly called

[1] https://github.com/delfimpandiani/visualsense. Access date: December 2023.

"word", and sentences to frames in a schematic structure based on the common scene they evoke. In FrameNet [270], a formal representation of Fillmore's frame semantics, frames are also explained as *situation types*. The Framester ontological hub [131, 128] provides a formal semantics structuring of commonsense knowledge in a curated linked data version from several, multimodal resources. The Framester ontology includes (besides FrameNet) linguistic resources such as WordNet [250], VerbNet [316]; a cognitive layer including MetaNet [130] and ImageSchemaNet [104]; and it is multilingual thanks to the alignment with BabelNet [262]. Furthermore, it includes factual knowledge bases (e.g. DBpedia [20], YAGO [343], etc.), and ontology schemas (e.g. DOLCE-Zero [134]), with formal links between them, resulting in a strongly connected RDF/OWL KG.

IV.1.2.3 ValueNet

The ontological module dedicated to formally representing moral and cultural values as semantic frames is ValueNet [103]. ValueNet formally represents three orders of values, according to the main theories in the literature. The first order (i) is composed of moral values, intended as universal, Kantian categories [193]. These values transcend the human species and are attested in the animal domain too [235]. They include values such as "Care", "Liberty" and "Equality". They are mainly modeled from Graham and Haidt's Moral Foundations Theory [148]. The second order of includes (ii) cultural values, whose existence is confirmed by several experiments [319, 320, 351], but their extensional semantics depend on sociocultural variables. The Basic Human Values theory by Shalom Schwartz [318, 145] models these kinds of values, such as "Openness to change", "Tradition", and "Self Enhancement". The ValueNet ontology includes also a module (iii) generated by those "values" gathered from non-official, non-specific, web-scraped concepts classified in several online repositories as "values". This bottom-up approach consists of a "return to the text": its aim is to include in the ontological representation those entities that, albeit not included in well-established theoretical frameworks, shape our daily behavior. These values are called "Folk values", and some examples could be "Punctuality", being punctual, "Intelligence", showing intelligence, and "Partnership", teaming up for a certain goal. The ValueNet repository is entirely available on the dedicated repository[2] and it is queryable from the Framester SPARQL endpoint.[3]

[2]The ValueNet GitHub repository is available here: https://github.com/StenDoipanni/ValueNet. Access date: December 2023.

[3]The Framester endpoint can be found here: http://etna.istc.cnr.it/framester2/sparql. Access date: December 2023.

IV.1.2.4 EmoNet

The ontological module dedicated to formally representing emotions as semantic frames is EmoNet. EmoNet in its current version includes the transposition of the Ortony, Clore, and Collins (OCC) appraisal theory [277], and the Basic Emotions (BE) theory by Ekman [114]. The emotions covered by this current version are the six Ekman Basic Emotions: Fear, Sadness, Anger, Enjoyment, Disgust, and Surprise. The EmoNet ontology is available online on its repository[4] and it is queryable from the Framester SPARQL endpoint.

IV.1.2.5 FRED

FRED[5] [132] is a hybrid statistical and rule-based knowledge extraction tool, able to generate RDF and OWL KGs taking as input directly text from natural language. Being directly linked to the Framester ontology it can be considered as a "situation analyzer". Its graphs include (i) word sense disambiguation to the WordNet resource, (ii) VerbNet verbs disambiguation, including the superimposition of VerbNet semantic roles attribution on the semantic arguments structure of the sentence; (iii) frame detection from FrameNet; (iv) PropBank frame recognition; (v) DBpedia entity linking. The usage of the FRED tool allows KG generation from natural language integrating the previously mentioned well-known semantic web resources aligned in the Framester hub while keeping the semantic dependencies structure of a sentence and a completely explainable knowledge enrichment pipeline.

IV.1.3 Approach

This section is focused on the VSO and VSKG development. Sec. IV.1.3.1 is focused on the original Visual Genome data model and the rationale for VG ontological transposition; Sec. IV.1.3.2 describes the enrichment of VG with semantic frames, values, and emotions; Sec. IV.1.3.2 is focused on populating the VSKG with the VG enriched data; and finally Sec. IV.1.3.2 describes how the VSO has been tested.

[4]The EmoNet GitHub repository is available here: https://github.com/StenDoipanni/EmoNet. Access date: December 2023.

[5]FRED online demo is available at http://wit.istc.cnr.it/stlab-tools/fred/demo/. Access date: December 2023.

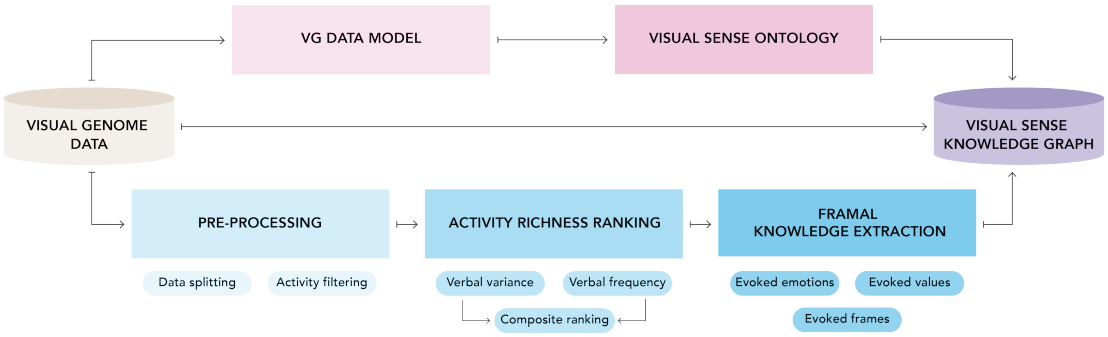


Figure IV.1.2: Starting from the data provided by the Visual Genome project (beige), our pipeline allows for the automatic creation of a semantic web KG containing visual, factual, and linguistic data. Top: Data modeling and ontology engineering branch (pink shades). Bottom: Framal knowledge enhancement branch (blue shades). The combination of the original data, the Visual Sense Ontology, and the extracted framal data are used to create the Visual Sense Knowledge Graph (purple).

IV.1.3.1 Data Modeling and Ontology Engineering Branch

In this section, we describe the ontological analysis of Visual Genome, as well as the rationale and modeling practices applied for its ontological transposition.

Visual Genome Data Model Analysis

The Visual Genome dataset can be accessed in two ways: through their API to directly access data from their server, or by downloading the entire dataset and working with it locally. In our case, we chose to download the data and parse it locally. Before designing our ontology, we manually explored the JSON files to better understand the structure of the data. During this exploration, we discovered issues such as the use of different keys for the same instances in different JSON files. To address this, we sketched out the implicit data model behind Visual Genome and proposed an intermediary data model to eliminate these duplications, and to streamline the ontology design process. We also consulted the provided documentation to gain a deeper understanding of the dataset's seven main components.

Visual Sense Ontology Engineering

eXtreme Design This project utilizes the eXtreme Design (XD) methodology [282] to develop the Visual Sense ontology, drawing inspiration from test-driven

practices in eXtreme Programming [35]. XD is based on the reuse of ontology design patterns (ODP), which address recurring modeling issues and are available in catalogs such as the Ontology Design Patterns Portal,[6] the Workshop on Ontology Design Patterns series,[7] and the University of Manchester catalogue.[8] In addition, XD adopts a modular design approach, which involves breaking down requirements into standalone, interchangeable modules. To collect requirements–Competency Questions (CQs) and General Constraints (GCs)–the methodology employs "stories," a set of sentences that illustrate the type of facts that the ontology should encode. To increase manageability, XD suggests dividing complex stories into smaller, more manageable ones that can be characterized by priority level, title, and identifier to indicate any possible dependencies on other stories.

Visual Sense: Stories and Competency Questions The XD methodology was applied to the Visual Genome dataset to comprehensively explore it and to guide the design of VSO. This involved formulating a general story that guides the exploration of the dataset, focusing on searching for images containing objects, identifying object attributes, discovering object relationships, analyzing object and relation regions, investigating conceptual frames, and exploring bounding box data in relation to frame evocation. The story also aims to identify synsets related to objects, regions, relations, and frames within the images. To support this exploration, the story was broken down into sub-stories, each with corresponding competency questions. In total, 26 competency questions were formulated that correspond to one or more sub-stories. The competency questions were developed using the XD methodology for ontology design, which involves identifying the regions where objects and relations are involved, their respective bounding boxes, and conceptual frames evoked by the images. The methodology also includes searching for synsets related to objects, regions, and relations. We report the sub-stories, their descriptions, and the corresponding competency questions (see Table IV.1.1). The complete document containing the competency questions can be accessed on the project Github.[9]

Besides the described stories and competency questions, the ontology and KG can be used to answer more complex and interesting queries, such as: What is the largest region (in terms of surface area) in an image, and what does it depict?

[6] http://ontologydesignpatterns.org/wiki/Main_Page. Access date: December 2023.

[7] http://ontologydesignpatterns.org/wiki/WOP:Main. Access date: December 2023.

[8] http://www.gong.manchester.ac.uk/odp/html/. Access date: December 2023.

[9] https://github.com/delfimpandiani/visualsense/blob/main/A-Data_Modeling_Ontology_Engineering_Branch/2_eXtreme_Design/competency_questions.md. Access date: December 2023.

Table IV.1.1: Substories and related Competency Questions (CQs) used as requirements for the design of the Visual Sense Ontology (VSO).

Story	Description	Related CQ
Substory #1	I want to search images that contain certain objects and want to find out which attributes are associated with them, alongside the images' path URLs.	CQ1, CQ21, CQ26
Substory #2	I want to know if there are relations in an image and, if this is the case, I want to identify the domain and range of each of the relations.	CQ13, CQ14, CQ15, CQ16
Substory #3	I am interested in identifying the image regions in which the objects and relations are involved, and their respective bounding boxes.	CQ2, CQ3, CQ8, CQ9, CQ10, CQ12
Substory #4	While searching for images that contain objects and relations, I want to know more details about the bounding boxes that cover these objects and relations, such as their location and size.	CQ3, CQ4, CQ5, CQ6, CQ7, CQ17, CQ18, CQ19, CQ20
Substory #5	I need to investigate if certain images evoke any conceptual frames, and in which regions these frames are involved in.	CQ22, CQ23
Substory #6	I want to search for synsets that are related to objects, regions and relations in a certain image.	CQ11, CQ24, CQ25

What are the coordinates of a depicted object in an image, and what percentage of the image does it occupy? Are there relationships between two objects of the same type in some image? Which images depict one human holding another human?

IV.1.3.2 Framal Knowledge Enhancement Branch

In this section, we describe the pipeline of knowledge enrichment from the Visual Genome dataset by reusing the Framester [131] ontological hub and the FRED tool.

Data Pre-processing

The original VG data required some pre-processing to be prepared for the frame evocation steps.

Data Splitting Due to the substantial size of the Visual Genome JSON files, we employed data pre-processing techniques as the first step, including parsing the scenegraph.json and regiongraph.json files provided by VG into splits of 10,000

image records each. The subsequent steps were designed to operate on each split individually, to facilitate the cleaning and filtering process.

Activity filtering To populate a KG with meaningful frame-related information from VG, we decided to focus on "action-oriented" images, where agents such as humans or animals are engaged in some kind of activity, and we used verbal relations as a proxy to identify actions. In VG, verbal relations are disambiguated to WordNet synsets (e.g., the *wearing* lexical unit in "a child is wearing a T-shirt" is disambiguated to `wn:wear-verb-1`, which in turn, in the Framester ontology, evokes the `fs:Clothing`, `fs:Dressing`, `fs:Wearing` frames). Therefore, we only considered images that are associated with at least one verbal relationship, assuming that such images would evoke more diverse frames and thus better enrich the KG. This decision is also reflected in the definition of `ImageRegion` class in the ontology T-Box (see Section IV.1.2.1), which must contain a verbal relationship between two objects in order to be included in the KG. To create a subset of action-oriented images, we iterated through the relations data provided by VG, tagging each relationship label with a part-of-speech (POS) tag, and selecting only those relationships that were tagged as verbs. We pruned out prepositional relations such as OF, ON, and WITH. We created a dictionary with image IDs as keys and the number of verbal relations for each image as values.

Activity Richness Ranking

To identify images that most likely contain useful linguistic information for detecting linguistic frames, we propose using an "activity richness ranking" approach. This approach measures the frequency and variety of verbs used to describe the actions taking place in an image. We used the ranking to determine the order in which to introduce images to the Visual Sense Knowledge Base.

Ranking Images by verbal variance This first criterion to rank images was based, in linguistic terms, on the number of verbal token occurrences per image. We calculated the total number of occurrences of verbal (non-prepositional) relationships labeled for each image. To visually explore the distribution of the ranked images, matplotlib[10] was used to plot the distribution of images according to the number of occurrences of verbal relations (Figure IV.1.5a).

Ranking images by verbal relation variance The second criterion was to rank images by their number of unique verbal types, i.e. by the number of unique verbal relationship types present in each image. Since conceptually we were more

[10] `https://matplotlib.org/`. Access date: December 2023.

interested in, for example, a scene with fewer occurrences of the same action but more types of different actions, a script was developed to count the amount of unique verbal relations per image. The results were plotted as well (Fig. IV.1.5b). Next, we calculate the "variety score" for each image, which is a measure of the diversity of verbs used in the image description. We do this by calculating the Shannon entropy of the verb frequency distribution for each image.

Composite Ranking We decided to do a final, composite ranking by normalizing the scores for each image on both criteria dimensions, computing the rankings for each dimension separately, and finally combining them into a single weighted average ranking. The weight for each dimension is set to 0.5, indicating equal importance.

Frame Knowledge Enrichment

From a single split, ranked by its activity richness, we operate a frame knowledge enrichment module according to the following steps: (i) region description retrieval, (ii) FRED KG generation from region description, and (iii) frame knowledge enrichment. The first step (i) consists in individuating all the regions per image in the considered split. Each region is described by a human annotator with a plain sentence (e.g. in Fig. IV.1.1 "the man is wearing a white t-shirt" or "the boy is on a skateboard"). Each of these sentences is then passed to the FRED tool (ii) to automatically generate a KG out of the description in natural language. Finally (iii), the FRED tool performs frame detection, and the FRED graphs are also used to query the ValueNet and EmoNet ontological modules to perform value and emotion detection. The final result consists of the enrichment of each region with FrameNet frames, ValueNet values and EmoNet emotions. This enrichment can be used to explore Visual Genome, exploiting the already existing entity and relation disambiguation to WordNet synsets which, in Framester, are aligned as frame evokers, improving, even more, the available querying material.

Visual Sense Knowledge Graph Creation

To populate the VSKG, we aligned the original VG data and extracted frame data with the VSO. First, we extracted relevant fields such as emotions, values, and frames from the TSV frame-related output of the knowledge enhancement branch to create a dictionary of frames for each selected image. Next, we loaded this data along with data from various VG sources. Initially, we attempted to perform the mapping using PyRML.[11] However, we found that an ad-hoc Python script utilizing the RDFlib library [206] provided greater agility and understanding.

[11] https://github.com/anuzzolese/pyrml. Access date: December 2023.

The final module imports various Python packages and defines functions for the string representation of synset and frame URIs to fit the Framester Hub syntax. Additionally, it defines a function that generates an image KG based on input data using the RDFLib Python library. The graph creation process utilizes several namespaces, including VisualSense (VS), the Common Procurement Vocabulary (CPV), Framester (FSCHEMA and FS), Haidt Values (MFT), Schwartz Values (BHV), Folk Values (FOLK), and Basic Emotions (BE). The scripts for this process are available in the GitHub repository.[12]

Visual Sense Ontology Testing

The eXtreme Design methodology emphasizes unit testing of the ontology, which involves Competency Question verification tests, Inference verification tests, and Error provocation tests [49]. These tests respectively validate whether the ontology can address the competency questions gathered during requirement collection, confirm that the inference mechanisms are established to ensure the proper execution of the inference requirements, and examine how the ontology behaves when given random or incorrect data. The competency question verification test consists of the reformulation of the competency question from natural language to SPARQL queries and running it against the ontology using a toy dataset which includes the expected result of the query. Inference verification tests are used to understand how the information needs to be produced, i.e. entered explicitly as assertions or derived from other facts through inferencing. Lastly, error provocation is a stress test of the ontology to verify how the ontology reacts when it is fed with erroneous facts or boundary data.

The XD methodology provides a thorough and descriptive protocol for the testing of ontologies. The competency questions collected for the development of the Visual Sense ontology have been tested with the Competency question verification test, while the requirements for the Inference Verification test and the Error Provocation test are found in the list below. The execution of each of the test cases is documented and its documentation includes the requirement that is being tested, the category of the test, the description of the test, the test itself, the input test data, the expected result, the actual result, the credentials of the tester, the test execution date, the execution environment, result, and comment. The protocol was followed partly manually and partly automatically with the help of the XDTesting tool,[13] an automation of the testing process based on the XD protocol and it is integrated with GitHub. The test cases, the datasets, and their

[12] https://github.com/delfimpandiani/visualsense/tree/main/C-Visual_Sense_Knowledge_Graph_Creation. Access date: December 2023.

[13] testing.extreme-design.info

documentation can be accessed in the GitHub repository.[14]

Inference verification test requirements include:

- Is `ImageBox` a subclass of `Region`? (In the ontology, `ImageBox` is a subclass of `SpaceRegion` and `SpaceRegion` is a subclass of `Region`.)
- Is `BoundingBox` a subclass of `Region`? (In the ontology, `BoundingBox` is a subclass of `SpaceRegion` and `SpaceRegion` is a subclass of `Region`.)
- If a class is the domain of the `:depictsDepictedObject` property, then it is a `DepictedRegion`.

Error provocation test requirements include:

- `ImageBox` and `ImageObject` are disjointed.
- `Object` and `Situation` are disjointed.
- `Object` and `Region` are disjointed.
- `Region` and `Situation` are disjointed.
- `DepictedObject` and `ObjectRelation` are disjointed.

IV.1.4 Results

IV.1.4.1 Data Modeling and Ontology Engineering Branch

Visual Genome Data Model Analysis

Analysis of Visual Genome dataset shows details about its components and ontological assumptions:

- **Region descriptions** are human-generated and localized in a region of an image with a bounding box. They are allowed to have a high degree of overlap with each other. Noun phrases in region descriptions are canonicalized to WordNet synsets.
- **Objects** are delineated by bounding boxes and canonicalized to WordNet synsets using a heuristic and 30 hand-crafted rules. The authors do not provide any explicit definition of what counts as objects, but their requirement to be covered by a bounding box points towards its semantics being of physical, depicted objects.

[14] https://github.com/delfimpandiani/visualsense/tree/main/D-Testing/6_Ontology_Testing. Access date: December 2023.

- **Attributes** predicate something about an Object, most commonly regarding color (e.g., yellow), states/ continuous actions (e.g., standing), sizes (e.g. tall), and materials (e.g. plastic). They are normalized based on morphology and mapped to WordNet adjectives using 15 hand-crafted rules. The most common attributes describing people are intransitive verbs describing their states of motion. Certain sports (e.g. skiing, surfboarding) are overrepresented due to an image bias towards these sports in the original image dataset.

- **Relationships** refer to connections between Objects, which are directed from a "subject" to an "object" and include actions (e.g. jumping over), spatial relations (e.g. is behind), descriptive verbs (e.g. wear), prepositions (e.g. with), comparative relations (e.g. taller than), and prepositional phrases (e.g. drive on). To canonicalize relationships, prepositions are ignored, and WordNet synsets are selected based on their sentence frames matching the context of the relationship. Root hypernyms of the verb-synset pairs are also considered to reduce noise, and 20 hand-mapped rules are included to correct for WordNet's lower representation of concrete or spatial senses.

- **Region graphs** are directed graph representations of a region, while **scene graphs** are the union of all region graphs for an image.

- **Question and answer (QA) pairs** include freeform and region-based QAs, and noun phrases in region descriptions are canonicalized to WordNet synsets.

Visual Sense Ontology

VSO is an ontology that aims to formally represent Visual Genome's annotation components and their interrelationships, and to connect these components to the Framester schema, so as to further ground visual data to language. The ontology was developed following the XD ontology design methodology, reusing ontology design patterns (ODPs) and aligning it to other ontologies (see the VSO T-Box IV.1.3). Below, we provide explanations of its crucial classes and properties. The Visual Sense ontology has been published at the following permanent IRI: `https://w3id.org/visualsense`.

VSO was aligned to DOLCE Ultra Lite (DUL)[15] foundational ontology due to the cognitive nature of VSO, as the task of representing and improving formal knowledge in the visual sense-making process is particularly coherent with

[15] `http://www.ontologydesignpatterns.org/ont/dul/DUL.owl`. Access date: December 2023.

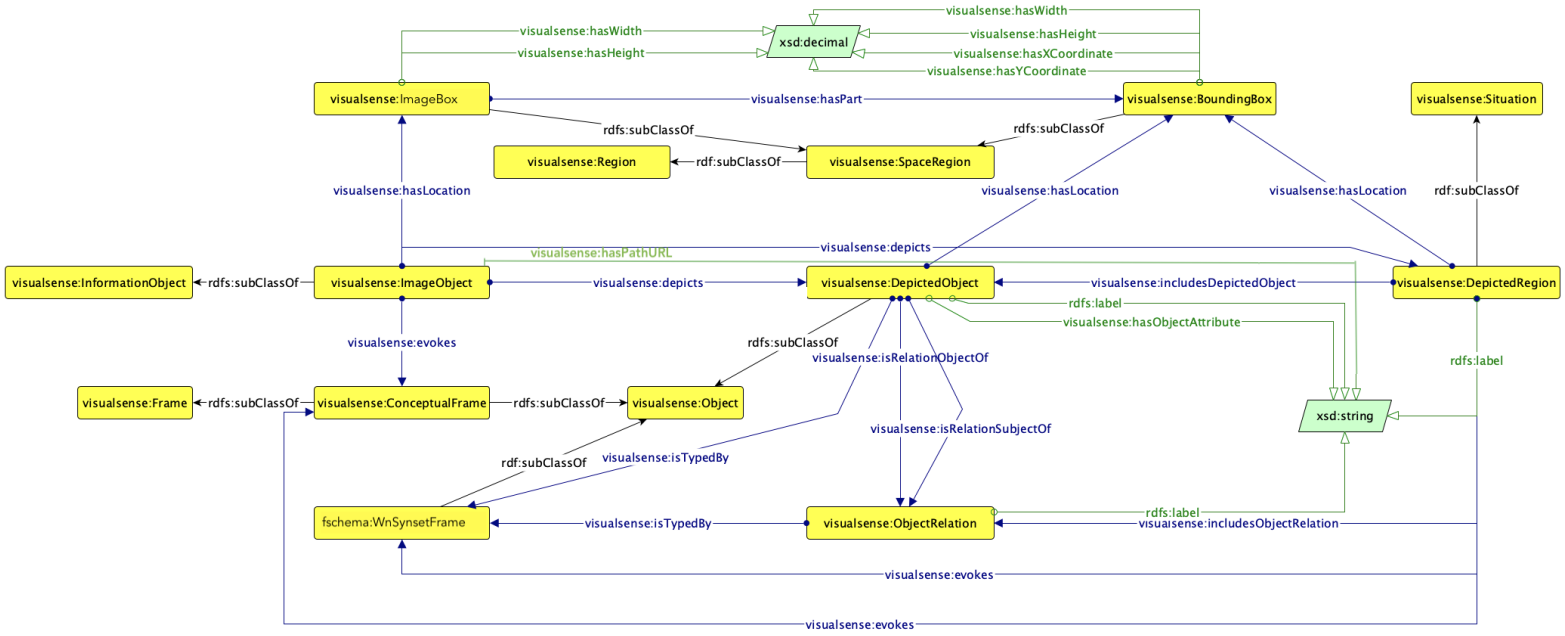


Figure IV.1.3: T-Box of the Visual Sense Ontology (VSO). It models images with a conceptual duality of images as information objects and as information realizations, and provides a framework for contextualizing depicted objects, relationships, and evoked conceptual frames within a depicted region.

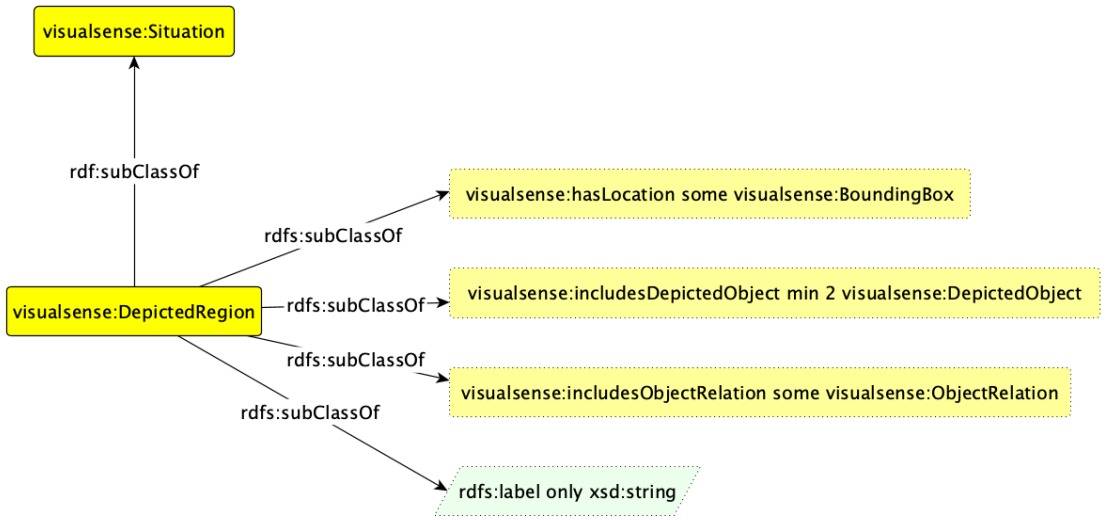


Figure IV.1.4: In VSO, a :DepictedRegion is modeled as a subclass of the class dul:Situation, in the sense that a depicted region provides a context and is the setting for a variety of things (depicted objects, relationships between depicted objects, evoked conceptual frames) that share the same informational space.

the human cognitive and socio-cultural aspects covered by DUL. What the Visual Genome model considers simply an "image", is considered in VSO as something that semantically is spread into two different classes, reusing the Content ODP Information Realization. `:ImageObject` is modeled as a subclass of `dul:InformationObject`, since the focus of expressiveness of this class is on the meaning that is conveyed in and by the Image as an object of information itself. This class of `:ImageObject` is furthermore axiomatized as having a realization through a location in some `:ImageBox`. The class `:ImageBox` is in fact a subclass of `dul:SpaceRegion` and it represents the physical extension of the image, the spatial area occupied by the image measured in terms of pixels.

This conceptual duality is coherently kept with all the other classes in VSO: a mereological relation exists between `:ImageBox` and any other subpart of the image, with the possibility to query the ontology based on the spatial area of interest. In particular, these physical subparts of an `:ImageBox` are the areas, bound by coordinates, which are recognised in VG as areas of location for Regions and Objects. They are modeled as instances of the class `:BoundingBox`, also subclasses of `dul:SpaceRegion`, and which are explicitly `dul:partOf` some `:ImageRegion`. They are also the `dul:locationOf` some `:DepictedObject` or of some `:DepictedRegion`.

The `:DepictedRegion` class applies the Situation Content Ontology Design

Pattern[16], whose intent is to represent contexts or situations, and the things that are contextualized. This pattern itself reifies the N-ary Relation Logical Ontology Design Pattern, and it allows the contextualization of things that have something in common, or are associated: a same place, time, view, causal link, systemic dependence, etc. In the case of VSO, `:DepictedRegion` is modeled as a subclass of the class `dul:Situation` in the sense that a depicted region provides a context and is the setting for a variety of things (depicted objects, relationships between depicted objects, evoked conceptual frames) that share a same informational space (Fig. IV.1.4).

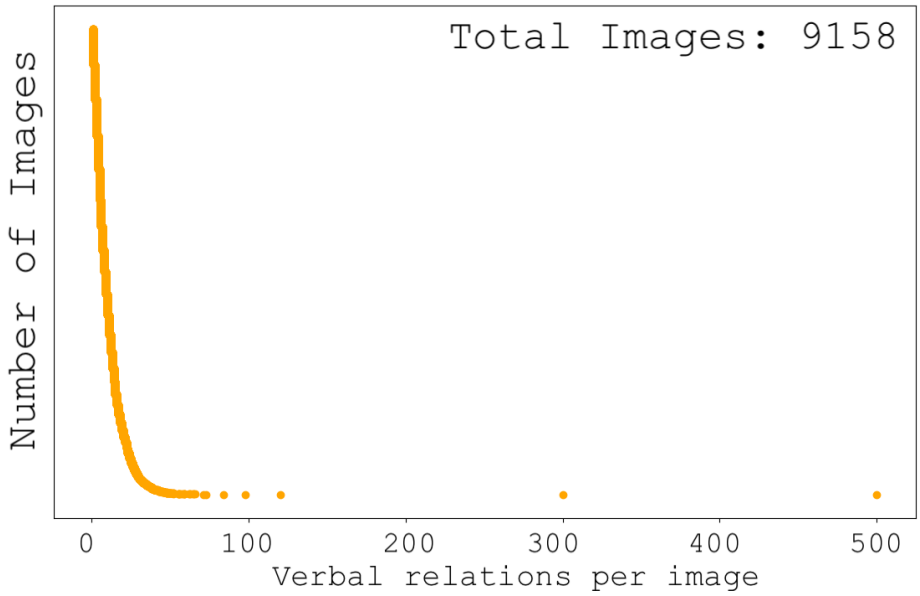
The other ontology reused in VSO is Framester schema, in particular the `fschema:Frame` and `fschema:ConceptualFrame` classes are reused both for the frames evoked by some `:DepictedObject`, located in some `:BoundingBox`, and the frames recognised as evoked by the FRED tool. Additionally, we reuse the `fschema:WnSynsetFrame` for the frames evoked by some specific Wordnet Synset.

IV.1.4.2 Frame Knowledge Enhancement Branch

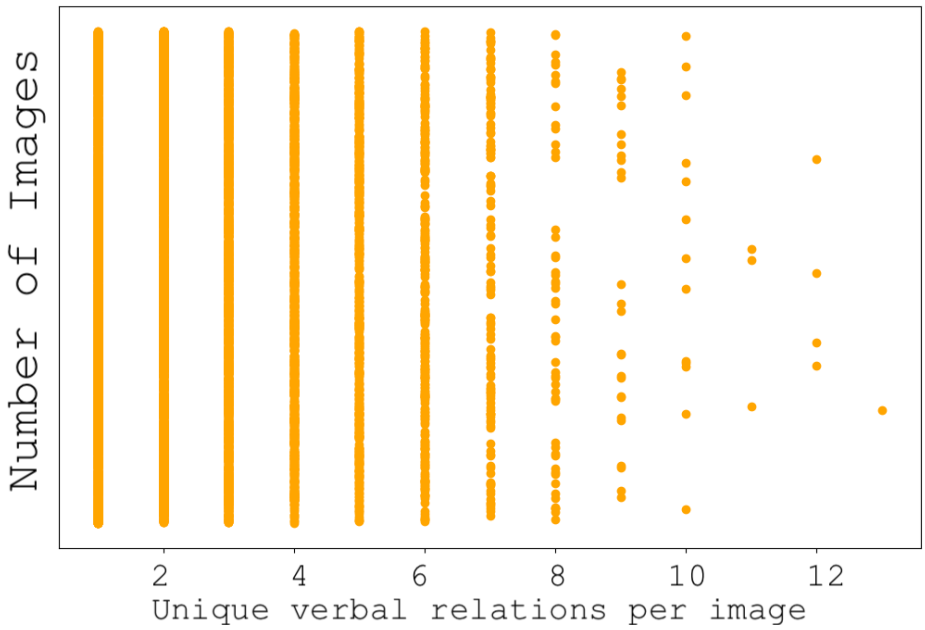
Activity Filtering and Ranking

In order to perform knowledge enrichment with frames, values, and emotions we applied the splitting and filtering methodology, ending up with 10 splits each with about 10k images. From here on the data shown refer to Split n. 2, spanning from object 10.000 to 19.999 in the original scene graphs VG data file. After applying the filter to retrieve only images with at least one verbal relation, the subset resulted in a total of 9158 images, meaning that over 90% of the images were labeled with at least one verbal relation. When we applied the first criterion of the activity richness ranking, we found that the vast majority of VG images lie in the span between 1 to 100 occurrences per image of verbal relations, with some peculiar graph outliers, namely the two dots showing images with about 300 and 500 verbal relation occurrences each (see Figure IV.1.5a). The application of the second criterion showed that the vast majority of images include only 1-7 unique verbal relations types (see Fig. IV.1.5b). We identified that the most common verbs were auxiliary verbs like “Have.v.1” and “Be.v.1”, while some other most commonly used verbs are those used to describe people in images, such as “Wear.v.1” or “Stand.v.1”.

[16] `http://ontologydesignpatterns.org/wiki/Submissions:Situation`. Access date: December 2023.



(a) Distribution of images in VG according to our first ranking criterion: the total number of occurrences of verbal relationships per image. The distribution follows a long tail distribution, with a big majority of images depicting little to no verbal relationships.



(b) The distribution of images in VG according to our second criterion: the number of unique verbal relationship types per image. The vast majority of images include only 1-4 unique verbal relations types.

Figure IV.1.5: The distribution of images in VG according to our two ranking criteria.

Framal Knowledge Enrichment

We analyzed the top 76 images from the second split using our frame extraction and value/emotion enrichment module, which yielded a total of 17,566 frame evocations. Of these evocation occurrences, 163 were identified as emotions from the EmoNet ontology. In terms of values, we found 1,410 occurrences of Folk values (socio-behavioral attitudes), 438 occurrences of cultural values from Schwartz's Basic Human Values, and 751 occurrences of moral values from Moral Foundations Theory. The remaining 14,804 evocations were identified as FrameNet frames.

IV.1.4.3 Visual Sense Knowledge Graph and Evaluation

The Visual Sense KG

Our pipeline automatically populated the KG with 76 images, resulting in more than 150,000 triples, which is available online following RDF standards.[17] The VSKG was generated using VG input about image metadata, image regions, and image scene graphs. Additionally, frame evocation data resulting from our framal knowledge enrichment branch was introduced, via the VSO. The pipeline automatically combined these three resources to represent images as information objects (`ImageObjects`), and their information realizations are represented as `ImageBoxes`. Conceptual frames are represented as Framester-based URIs within the KG. VSKG instantiates all classes defined by the VSO, capturing rich knowledge about the 76 images, including co-occurring depicted objects, depicted relationships, attributes, pixel-based locations and dimensions, evoked WordNet synsets, and Conceptual Frames evoked by specific regions.

Testing and Evaluation

For the testing of the Visual Sense ontology, we have created 34 test cases, out of which 26 are Competency Question verification tests, 3 are Inference Verification test, and 5 are Error Provocation tests. Each of the test cases was executed and their results were documented. From the 26 Competency Question Verification tests, 25 passed successfully, from 3 Inference Verification tests, 2 passed successfully, and from 5 Error Provocation tests, 0 passed successfully. The result of only one Competency Question Verification test was not successful, and it was caused by a syntax error. The Inference Verification test that resulted in failure is caused because the requirement cannot be inferred by the ontology. Lastly, the Error Provocation tests were not passed because the ontology did not include disjoint axioms

[17] https://github.com/delfimpandiani/visualsense/blob/main/C-Visual_Sense_Knowledge_Graph_Creation/4_KG_Generation/output_KG/visualsense.ttl. Access date: December 2023.

between the concepts `ImageBox` and `ImageObject`, `Object` and `Situation`, `Object` and `Region`, `Region` and `Situation`, and `DepictedObject` and `ObjectRelation`. The test results (especially the failures) bring to attention requirements that need to be developed into the ontology. In particular, the absence of the disjointness axioms is consistent with the formal representation of social objects, and is resolved through the import of the DOLCE foundational ontology, and the reuse of the ontology design pattern Descriptions and Situations (DnS).

IV.1.5 Discussion

The primary objectives of this study were to: (i) enhance the Framester resource by expanding its coverage in a multimodal direction through the integration of the VG annotated dataset; (ii) make the VG dataset directly and explicitly queryable by transposing its data model into RDF linked open data format; and (iii) enrich the visual data with knowledge from several layers (including moral/cultural values, emotions, and other conceptual frames) to align it with the factual and linguistic knowledge already present in the Framester Hub. This alignment is essential to address state-of-the-art problems, such as commonsense knowledge and multimodal extraction tasks, which can benefit significantly from the Visual Sense Ontology and the link between frame evocation triggers and situational occurrences IV.1.6. Thus, this work makes a valuable contribution to the field of knowledge representation and semantic web technology. The ensuing paragraphs delineate the primary achievements of this project.

Activity-Focused VG Preprocessing Pipeline A full preprocessing pipeline of Visual Genome data is available on the Visual Sense GitHub repository that allows for preprocessing for action- or frame-focused tasks. It includes data trimming into more tractable splits, as well as data filtering.

Image Activity Ranking The methodology described in Sec. IV.1.3.2 to process images according to the "richness of depicted activity" measured by the presence, amount, and variance of verbal relations is a useful method to establish an activity richness ranking, and can be useful to select the most interesting images for detecting semantic frames. In fact, by measuring both the frequency and variety of verbs used in image descriptions, we create a composite ranking that is more informative than either dimension alone.

Visual Sense Ontology The Visual Sense ontology is a significant achievement in knowledge representation and semantic web technology. This ontology is the

product of a reverse engineering process that extracted the conceptual model underlying the Visual Genome dataset. It was designed to align with the DOLCE Ultralight foundational ontology and is integrated into the Framester schema using best practices in ontology modeling. These practices include reusing ontology design patterns, aligning entities, axiomatizing classes, and conceptually grounding with the DnS pattern. Moreover, our approach extends beyond previous efforts by conceptualizing images as information objects and information realizations in relation to linguistic frames. This perspective allows for a more comprehensive, accurate, and accessible way of integrating and analyzing multimodal data. By adopting this approach, VSO provides a robust framework for tackling complex challenges, such as multimodal extraction tasks and commonsense knowledge.

Framal Knowledge Enrichment This study presents a comprehensive pipeline for frame evocation from visual data descriptors, which is capable of tackling the problem of corrupted or incomplete data through the reuse of the FRED tool and WordNet lexical unit disambiguation. The pipeline takes as input the region and scene data provided by Visual Genome and generates a complete list of frames evoked and the number of evocations per image. Importantly, the framal knowledge enrichment module is able to detect the evocation of not only FrameNet frames, but also ValueNet values and EmoNet emotions.

Visual Sense Knowledge Graph Population In addition to the ontological structure, this study offers a pipeline for populating the ontology with data from both the original VG files and frame evocation. This pipeline is capable of generating a queryable KG for any image in the original VG dataset. Overall, this pipeline enables the systematic and efficient population of the ontology with data, The resulting image KGs provide a rich resource for researchers and practitioners to analyze multimodal data and extract valuable insights.

IV.1.5.1 VSKG and its Potential Uses

SPARQL Querying Visual Genome This work offers a significant contribution by enabling direct and explicit querying of the VG dataset. Through transposing VG's data and model into RDF-linked open data format, a resource is created that enables complex queries, particularly SPARQL queries, to retrieve knowledge about various aspects of visual content, including entities and their attributes, regions in images, verbal relationships among entities, entity-to-entity relations, the presence of a specific WordNet synset, frame evocation, and the presence of certain emotions or moral/cultural values. Additionally, using the Framester structure, it is possible to query for images that contain hypernyms of a specific entity (ac-

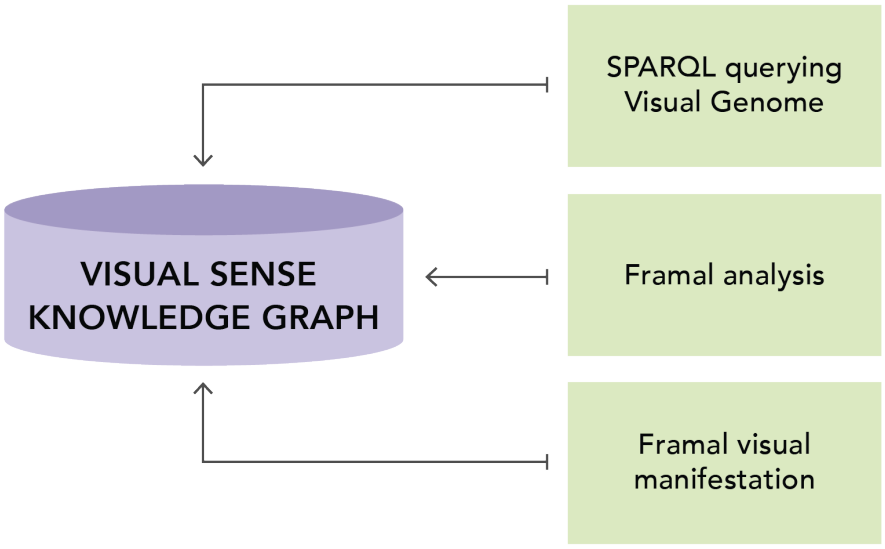


Figure IV.1.6: The VSKG is a versatile multimodal resource that offers multiple benefits. Not only does it store multimodal information pertaining to VG images, but it also enables direct and explicit queries on relationships within and between images and their contents. As a result, it facilitates sophisticated frame analysis of images, as well as exploration of patterns of frame compositionality. Furthermore, the VSKG's inclusion of both linguistic frame information and precise evocation coordinates allows for literal visualization of linguistic frame visual manifestations.

cording to WordNet hyperonymy relation), WordNet synsets evoking a specific frame, and entities with certain `fschema:SemanticType` such as, for example, generic containers like boxes and bottles. Queries can also be done to learn about the positionality of objects, for example, extracting data and statistics of the average distance between objects in specific relationships (e.g., measuring overlap or calculating the distance between the center of the bounding boxes of the object and subject). To demonstrate the querying capabilities, several SPARQL queries are available on the Visualsense GitHub repository.

Frame Detection Analysis Critically, the frame evocation is performed to be localized in specific regions of an image. Those same regions, in VSKG are connected to depicted objects and relationships these may be in. As such, VSKG provides an opportunity for deep exploration of visual content and framal analysis, enabling researchers to retrieve images based on complex relationships between frames. For example, one can retrieve all images featuring a region evoking the `fs:Animals` frame, where an object relationship evokes the `be:Enjoyment` emotion, and a region evoking the `fs:People` frame to retrieve images of some-

Figure IV.1.7: Four examples of framal visual manifestations on images from the VG dataset, with visible bounding boxes of depicted regions labeled with evoked frames, emotions, and values. Clockwise: the first image shows how frames that refer to concrete entities (fs:Clothing) can co-activate with more abstract frames such as (fs:Temperature) in the same image region. The second image provides visual instantiations of general frames like fs:CommerceScenario and folk:BodyMovement. The third image shows visual instantiations of frames like fs:Electricity around the electric guitar, and folk:PerceptionActive in the area of spectators paying attention to the performers. The last image demonstrates a visual instantiation of the value of folk:Partnership, among others.

one playing with a pet or animal. Moreover, the N-ary relation structure of frames enables automatic retrieval of semantic roles that participate in certain events by reusing VerbNet roles such as Agent and Patient/Undergoer. These roles are expressed through the VG Subject and Object annotations of a relation, respectively. Complex SPARQL queries are available in the Visual Sense Ontology repository and can be executed at the Framester SPARQL endpoint.

Framal Visual Manifestations The frame evocation pipeline we propose is an explainable process that leverages entities from open linked data and connects

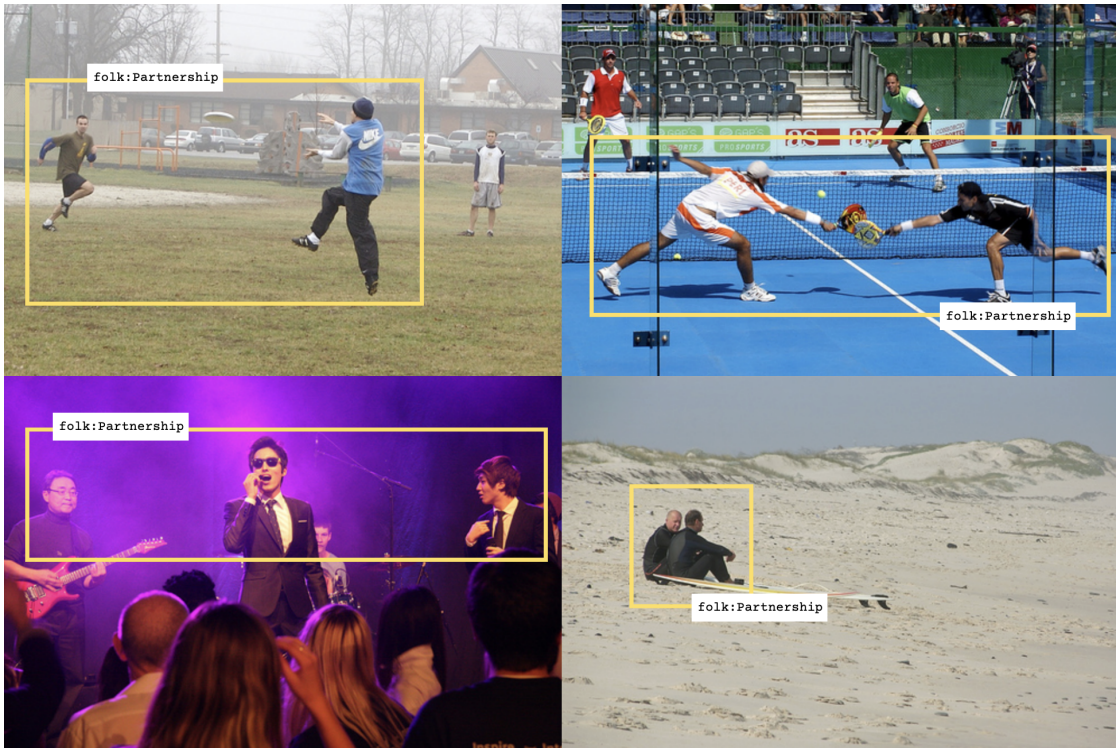


Figure IV.1.8: Four examples of very different folk:Partnership evocation. The visual instantiation of this frame shows how high-level semantics can be conveyed as commonsense knowledge through the same concept. In clockwise order: the first image shows young men coordinating in a motor activity while playing. The second image shows an official doubles tennis match. The third image shows two people chatting in the sand, with a more emotional and empathy-tinged nuance of partnership. The fourth shows a band performing on stage in front of an audience.

visual material to well-known semantic web resources. This pipeline not only enriches visual material with frame evocation but also enables the opposite flow of information: the retrieval of visual instantiations of semantic frames. For instance, the Visual Sense Ontology allows for the retrieval of all images evoking the `fs:Food` frame. The first image in Figure IV.1.7 demonstrates how frames that refer to more concrete entities, such as `fs:Clothing`, can be co-activated with frames like `fs:Temperature` in the same image region. Moreover, Fig. IV.1.8 shows how visual instantiations of values such as `folk:Partnership` can be explored to see how the same concept can be manifest in various types of situations.

IV.1.5.2 Future Work and Research Directions

There is a need for further improvements to enhance the smoothness of each pipeline's part. Despite the concreteness of the results, the ideal pipeline should generate both frame evocation and KG from the input of image ID alone after data polishing. However, due to the vastness of data and the requirement to iterate through each split of the original files, this passage still requires manual input to the specific previously extracted file from which to generate both frame evocation and KG. In light of this, some goals for the next iteration of this project are further expansion of VSO and VSKG, as well as the execution of further alignments and inferences. Further directions for the expansion of VSO and VSKG include:

- **Action Relevance Refinement**: A further step could focus on what VG refers to as "attributes". While most attributes in VG refer to physical qualities of material objects, such as color or texture, some attributes refer to "states" (e.g., "skiing", "sitting"). Modeling these attributes in a more sophisticated way could lead to more semantically complex KGs, allowing for more complex queries to be performed. The work could also benefit from more sophisticated activity richness ranking, by also taking into account state-of-the-art metrics of visual interestingness [85].

- **Prepositional Knowledge Integration**: VSO and VSKG do not currently include a vast part of the Visual Genome original dataset, specifically all the prepositional relations. Integrating prepositional knowledge would allow for the modeling of not only the framality layer of images but also the image-schematic layer of visual knowledge.

- **Integration of Question-Answer Pairs**: VSKG does not currently incorporate data related to Visual Genome's QA pairs, which are also disambiguated with synsets. These could be extracted and used to further refine or improve frame evocation.

- **Compositionality of Frame Evocation**: Further inference could be made on the compositionality of frame evocation, particularly in complex scenarios that represent abstract concepts like "violence" and "social disorder," or feelings like "love" and "shame."

- **Relevance of Frame Activation**: Further inferences can be performed about the relevance of some frame activation depending on the dimension of its region's box in comparison to the total surface of the image it belongs to.

- **Further Frame Roles Localization**: Considering a frame evocation within a specific `DepictedRegion` that contains an `ObjectRelation`, we also

have information about the Subject and Object of this relation from the Visual Genome dataset. A further direction is to align the Subject and Object with the `framesterrole:Agent` and `framesterrole:Undergoer` roles, respectively. This would enable defining the `fschema:TropeType` for both Subject and Object synsets, categorizing them as `TropeRoles` associated with the evoked frame.

- **Sophisticated Querying**: More sophisticated SPARQL queries could lead to further discoveries in the dataset. The existing competency questions tested are simple and straightforward regarding the built ontology, but there is immense potential for discovering knowledge with more complex competency questions.

IV.1.6 Conclusions

Guided by the overarching research question of whether visual data descriptors can effectively bridge the gap between raw pixel data and ACs, this chapter explored the feasibility of automatically reasoning over the concrete semantics of visual data to establish the evocation of high-level frames. The hypothesis that interpretable connections between images and high-level conceptual frames can be achieved through leveraging background commonsense knowledge and ontology-based automatic reasoning on concrete descriptors was affirmed. This chapter marks substantial progress in automating multimodal sensemaking by linking linguistic frames to their corresponding visual manifestations using ontology-based knowledge engineering techniques. The work has successfully integrated the extensively annotated dataset, Visual Genome, [233] within the Framester hub [131], extending the already comprehensive Framester resource coverage into the multimodal domain. Through the development of VSO, a pipeline for region-specific frame, emotion, and value evocation, and the construction of the VSKG, the project has effectively mapped data from its original source to create a queryable KG. This work signifies a significant advancement in the automation of frame evocation and multimodal sensemaking, with the potential for application across various fields.

Chapter IV.2

Situated Ground Truths: Bias-Aware AI with SituAnnotate

Summary One of the primary challenges in AC image classification is the subjectivity and cultural variability in image interpretation. Our research in Chapters II.2, III.1, and IV.1 has shown that concrete visual elements in images can effectively function as cognitive intermediaries to help bridge the semantic gap between raw pixel data and abstract concepts. However, these approaches rely on the initial assignment of labels to images, which can be a source of subjective bias itself. This process mirrors the current AI landscape, where annotations in the form of words or labels are pivotal for training AI systems. Notably, these annotations often lack essential contextual information, which can introduce biases. To address this challenge, we propose SituAnnotate, a novel ontology designed for "situated grounding," anchoring ground truth data in their contextual and culturally-bound origins. SituAnnotate provides a structured and context-aware data annotation approach, addressing potential bias issues with isolated annotations. It encompasses situational context, including annotator details, timing, location, remuneration schemes, annotation roles, and more, ensuring semantic richness. Aligned with the foundational Dolce Ultralight ontology, it offers a robust and consistent knowledge representation framework. Our approach produces structured, machine-readable knowledge that reduces subjectivity and cultural bias in AI systems by considering contextual annotation factors. As a tool for creating, querying, and comparing label-based datasets, SituAnnotate enables AI systems to undergo training with explicit consideration of context and cultural bias. This enhances system interpretability and adaptability, enabling AI models to align with diverse cultural contexts and viewpoints.

IV.2.1 Introduction and Background

In J. L. Borges' famous essay *The Analytical Language of John Wilkins* [51], animals are classified into unconventional and seemingly bizarre categories.[1] The story showcases the arbitrary and culturally-specific nature of categorization, a philosophical questioning into the complexities and subjectivity inherent in the act of classification. This theme finds a modern parallel in the rapid growth of artificial intelligence (AI) and data-driven applications, where classifying data is essential for training our machines [283, 34, 313].

Labeled data, which underpins modern AI systems, is the result of vast processes of data annotation, where meaning is assigned most commonly through linguistic labels to data points. Given that data produced and annotated by humans possesses unique value, with the underlying belief that the "human touch" is indispensable to ensure accuracy and quality, the annotation process often depends on microlabor of human platform workers [360]. Data annotation is deceptively complex, revealing a paradox where seemingly objective AI systems grapple with subjective annotations, resulting in inherent bias. This stems from the context-dependent nature of annotation, which challenges the notion of universal objectivity. In the digital age, AI systems, portrayed as objective, are constructed using data steeped in the subjectivity they aim to overcome. This intricate interplay between classification, subjectivity, AI data labeling, and bias emphasizes the complexities of modern AI development.

Data labeling processes are frequently shaped by human judgments, cultural viewpoints, and personal biases. It's important to clarify that the biases discussed in this work should not be conflated with the "bias" term in machine learning models, which, mathematically speaking is an intercept or offset from an origin. Rather, we are focusing on cultural bias in the sense defined by [353]:

> *the tendency to interpret and judge phenomena in terms of the distinctive values, beliefs, and other characteristics of the society or community to which one belongs.*

This chapter delves into the technical aspects of accounting for cultural bias in the process of assigning semantic labels to data, with a case study of how this bias permeates the moment of labeling pixel areas of images within training datasets. This particular phase of human-led or human-evaluated annotation is critical, as

[1] Borges' story presents a fictitious taxonomy of animals, supposedly taken from an ancient encyclopedia, which divides all animals into " (a) those that belong to the emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel hair's brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance."

the labels generated here become integral parts of input data for widely utilized models across various domains. Consequently, the "data itself" can harbor human biases, including stereotypes, prejudice, and racism. In this sense, this work primarily concerns itself with the intersection of cultural and measurement bias, with measurement bias denoting faulty, low-quality, or unreliable data collection measures, which can have many causes such as insufficient label options (e.g. binary gender [315]) or from subjective views from labelers. These biases can lead to skewed interpretations and annotations, subsequently affecting the decisions made by AI systems. A compelling example of this phenomenon can be observed in the realm of computer vision, where popular datasets like ImageNet [105] play a pivotal role by providing ground truths or "factual" meanings to extensive image collections. Paradoxically, these datasets inadvertently reinforce contested political categories and cultural prejudices. For instance, an image of an indigenous person in traditional attire might be labeled as "half-naked," perpetuating a culturally biased perception as objective truth. Classification frameworks hold authority in determining the significance of features, potentially amplifying specific worldviews while marginalizing others. Consequently, the ramifications extend beyond mere representation, encompassing the ability to mold societal perspectives and fortify preexisting biases.

The sway of these data biases is not limited to equity or fairness; it can significantly shape the very performance of AI systems reliant on them (e.g., [99, 384]). Over the past decade, the issue of data bias has taken center stage [269, 95], with endeavors to "unbias" models and/or the data that they learn from have become a cornerstone in the pursuit of equitable AI systems [75]. However, any effort to encapsulate the intricate realities of the world inherently carries with it biases and perspectives rooted in context. In this sense, rather than the pursuit of defining and cultivating "unbiased" datasets—an increasingly improbable feat—a paradigm shift is emerging, which uses biased datasets with the awareness of this phenomenon, and tries to identify how bias affects results, embracing the nuanced, situated nature of annotations [10, 301, 383]. It is in this context that we propose the SituAnnotate ontology, a knowledge representation, and capture tool poised to navigate the landscape of annotation situations for labeling data used to train AI systems.

IV.2.1.1 SituAnnotate to Enhance Cultural Bias-Aware AI

While considerable effort has been invested in establishing standards for capturing metadata pertaining to data and model production and reuse (e.g., data sheets [136] and model cards [252]), there is a lack of technical tools that allow both humans and machines to reason over such contextual information, a gap that persists especially at the level of singular annotations. this work advocates for

the explicit encoding of situational metadata alongside annotated data, to allow reasoning. This encoding should be designed to be both machine-readable and comprehensible/retrievable by human users.

this work introduces SituAnnotate, an ontology-based module designed to formally represent the culturally-bound processes involved in annotating data. It builds upon the Description and Situations ontology design pattern [129] to account for two key aspects: 1) the explicit tracking of culturally coded annotation situations, detailing how meaning is associated with data, and 2) the ability to reason over and compare annotations and their contexts. SituAnnotate, offers a structured and context-aware approach to annotating situational context, encompassing annotator type, temporal and spatial information, remuneration schemes, annotation roles, and more. SituAnnotate's core objective is to capture the contextual backdrop surrounding annotations while providing machine-readable representations of the circumstances in which data gains significance through linguistic labels. It builds upon the Dolce Ultralight ontology, ensuring robustness and consistency in knowledge representation, thereby facilitating the selection of specific data subsets based on annotation context criteria.

IV.2.1.2 Structure of the Chapter

This work is structured as follows: In Section IV.2.1.2, a review of related works is presented, covering AI data labeling practices, biases, and existing approaches to mitigate them. Section IV.2.2 introduces the SituAnnotate ontology, first describing the user requirement scenarios that guided the design of the ontology, and then defining fundamental concepts and design principles, and describing the core Classes. Section IV.2.3 discusses the case study of image annotations within computer vision pipelines. The evaluation protocol, including competency questions and results, are discussed in Section IV.2.4 The implications, contributions, and an example of module specialization of SituAnnotate are discussed in Section IV.2.5. Ultimately, Section IV.2.6 provides a concluding segment summarizing the key findings and the impact of SituAnnotate. The ontology is available online[2] and documented in its GitHub repository.[3] The latter also contains the SPARQL queries and tests used for the evaluation of the ontology.

[2] https://w3id.org/situannotate. Access date: December 2023.
[3] https://github.com/delfimpandiani/situAnnotate/. Access date: December 2023.

Related Work

IV.2.1.3 Annotated Data Hunger

The significance of data in the realm of machine learning cannot be overstated. As [307] succinctly puts it, "ML is data-hungry. Deep learning is data-ravenous." To effectively train supervised models, datasets with meticulously annotated labels are imperative, as they furnish the necessary supervised information to guide model training and estimate functions or conditional distributions over target variables from input data. Nevertheless, the process of manually labeling data can be labor-intensive and time-consuming. In response to this challenge, there are alternatives such as pseudo-labeling and label propagation, as discussed by [375], which offer the possibility of automatically annotating extensive unlabelled datasets based on a limited set of accurate annotations. This process then makes available ground truths an indispensable foundation for reliable model performance assessment and validation.

IV.2.1.4 The Human Touch in Annotated Data

Data annotation, as highlighted by [360], predominantly relies on human involvement, recognizing the unique value attributed to data produced and annotated by humans. This underscores the crucial role of the "human touch" in ensuring the accuracy and quality of annotated data. Geiger et al.'s work [137, 138] offers a comprehensive review of the landscape of human labeling of training data in machine learning, delving into best practices in this field. They argue that much of this labeling work aligns with structured content analysis, a methodology supposed to be "systematic and replicable" [296, p. 19] and historically employed in the humanities and social sciences to transform qualitative or unstructured data into categorical or quantitative data. This structured content analysis entails the work of "coders" or "labelers" who individually assign labels or annotations to items in the dataset according to "coding schemes", after which inter-rater reliability is assessed. Historically undertaken by students, crowdwork platforms like Amazon Mechanical Turk have become most common for data labeling tasks, with new platforms emerging to support micro-level labeling and annotation, including, for example, citizen science initiatives where volunteers collaborate to label data across various domains (e.g., [73]).

IV.2.1.5 The Garbage In, Garbage Out Principle

In the realm of machine learning, the axiom "garbage in, garbage out" [24] reverberates as a familiar cliche, emphasizing that the quality of data used in a process

directly influences the quality of the outcomes. Garbage data extends to include not only inaccuracies but also decontextualized or biased information that lacks relevant connections or meaning. Data quality concerns are often overlooked in ML research and education [138], but it is essential for those applying ML in real-world domains to grasp the implications of low-quality or biased training data. The idea that automated systems are not inherently neutral and instead reflect the priorities, preferences, and prejudices of those who have the power to mold artificial intelligence is an increasingly public topic of discussion, especially given that many datasets are systematically biased along various axes, including race and gender, which impacts the accuracy of those ML model. For example, [61] investigates the false assumption of machine neutrality, and the *coded gaze*–the algorithmic 'way of seeing' which classifies content through researcher- and machine-labeled categories–which "reflects both our aspirations and our limitations" [62, p. 44]. Another example is how the geographical sampling of Flickr images as well as the use of English as the primary language for dataset construction and taxonomy definition result in inherent cultural bias within the datasets [373], with work being done to design new annotation procedures that enable fairness analysis [317]. As such, evaluating supervised models solely with a held-out subset of the training data can obscure systematic flaws, especially in cases where the model is used for contentious decisions like those in finance, hiring, welfare, and criminal justice.

IV.2.1.6 Identifying and Documenting Bias in Data

AI research often relies on biased perspectives in ground truth datasets, potentially causing issues when lacking proper context. New frameworks aim to clarify the assumed knowledge within datasets and deployed AI systems to combat this problem.

De-biasing ML

There are efforts to "de-bias" ML (surveys by [248, 125]), including via developing domain-independent fairness metrics to test and modify trained models or predictions. For example, [142] addresses the issue of social biases in AI algorithms by proposing D-BIAS. D-BIAS is a visual interactive tool that employs a human-in-the-loop AI approach to audit and mitigate social biases from tabular datasets. D-BIAS uses graphical causal models to represent relationships among features in the dataset and inject domain knowledge. Users can detect bias against specific groups, such as females or black females, and refine causal models to mitigate bias while minimizing data distortion. Other approaches have been through dataset preprocessing [66] or database repair [312].

Documenting (Meta)Data

Other efforts have designed standards for capturing metadata pertaining to data and model production and reuse. [252] propose the use of "model cards" to accompany trained machine learning models, which are concise documents that provide benchmarked evaluations of models under various conditions. These cards also disclose the intended use cases, evaluation procedures, and relevant information about the model. [136] introduce the concept of "datasheets for datasets" drawing an analogy to datasheets for electronic components. They propose that every dataset should be accompanied by a datasheet that documents its motivation, composition, collection process, recommended uses, and more. This approach facilitates better communication between dataset creators and consumers, prioritizing transparency in data collection. Other approaches include "data statements" [41], "nutrition labels" [169], a "bill of materials" [28], "data labels" [43] and "supplier declarations of conformity" [16]. Additionally, [184] argue for the importance of a new specialization within machine learning focused on methodologies for data collection and annotation. They draw parallels with archival practices, where scholars have developed frameworks and procedures to address challenges like consent, power, inclusivity, transparency, ethics, and privacy. By incorporating these approaches from archival sciences, they encourage the machine learning community to be more systematic and cognizant of data collection, particularly in sociocultural contexts.

Investigating Annotator Bias

Moreover, efforts to enhance transparency and accountability in the ML community have focused on detecting and addressing annotator bias. [384] identify annotation bias by analyzing similarities in annotator behavior. To achieve this, they construct a graph based on annotations from different annotators, apply a community detection algorithm to group annotators, and train classifiers for each group to compare their performances. This approach enables the identification of annotator bias within a dataset, ultimately contributing to the development of fairer and more reliable hate speech classification models. Within the context of hate speech detection systems, [10] delves into the issue of annotator bias with a specific focus on demographic characteristics. They construct a graph based on annotations from various annotators and utilize community detection algorithms to group annotators based on demographics. They then proceed to train classifiers for each demographic group and conduct performance comparisons. This rigorous approach enables them to shed light on how demographic features like first language, age, and education significantly correlate with performance disparities.

IV.2.1.7 Ontologies for Digital Hermeneutics

Ontologies formally represent data semantics in a machine-readable format, enabling explicit semantics and facilitating queries based on concepts and relationships [27]. Previous research has applied ontology-driven approaches in fields like image understanding and computer vision, especially in addressing the challenge of image interpretation. A recent work focuses on modeling interpretation and meaning for art pieces, presenting a data model for describing iconology and iconography. Additionally, the Historical Context Ontology (HiCO) aims to outline relevant issues related to the workflow for stating and formalizing authoritative assertions about context information for cultural heritage artifacts [96]. Also, the VIR (Visual Representation) ontology, constructed as an extension of CIDOC-CRM, sustains the recording of statements about the different structural units and relationships of a visual representation, differentiating between object and interpretative act [68]. These developments illustrate the versatility of ontologies in addressing various interpretation challenges in different domains.

IV.2.2 The SituAnnotate Approach

IV.2.2.1 Situating (Ground) Truths

this work contends that a crucial step towards the goal of responsible and ethical AI [87] involves the deliberate grounding of assumed objective truths within their respective situated contexts. This view aligns with the growing need for technical solutions to challenge the conventional notion of an unequivocal truth in human annotation [19], to adopt a power-aware approach to data design and production [249], and to reveal how AI, ML, and data practices inadvertently perpetuate colonial power dynamics and value systems [48, 256].

We philosophically adhere to the idea of "situated grounding" in training data, echoing the concept of "situatedness" exemplified by Donna Haraway in 1988. Haraway challenges the traditional detached view of vision, characterized as a "conquering gaze from nowhere":

> *This is the gaze that mythically inscribes all the marked bodies, that makes the unmarked category claim the power to see and not be seen, to represent while escaping representation.* [158, p. 581]

We are inspired by Haraway's alternative paradigm of "situated knowledges," advocating for a perspective rooted in complex, contradictory, structured bodily experiences, rather than from an assumed objective standpoint [158, p. 589]. Embracing this paradigm involves recognizing the multifaceted nature of localized knowledge.

IV.2.2.2 Scenarios

Our ontology, SituAnnotate, aligns with Donna Haraway's 'situated knowledges' paradigm, emphasizing context-dependent perspectives over detached objectivity. To ensure its effectiveness, we devised 11 user requirement scenarios, serving as practical examples of the intricate challenges SituAnnotate addresses. These scenarios highlight that ground truths are context-dependent, nuanced entities. In this section, we present these scenarios as practical use cases, showcasing how SituAnnotate provides valuable insights and supports various annotation-related tasks.

Scenario 1: Geographic Distribution of Annotation Situations
I want to understand the geographic distribution of annotation situations in SituAnnotate. Specifically, I want to know which countries have been the location of annotation situations, how many annotation situations were located in each country, and which country has the highest number of annotation situations.
Rationale: This scenario aims to shed light on the geographic scope of annotation situations captured by SituAnnotate. Understanding where annotation activities are concentrated can provide insights into regional preferences, data availability, and potential biases in the annotation process.

Scenario 2: Temporal Filtering of Annotation Situations
I want to research the temporal aspects of annotation situations. Specifically, I want to select a specific period of time and identify which annotation situations a particular image has been involved in during that time. This allows me to track the history of annotations for the image and observe how they may evolve over time.
Rationale: This scenario tests SituAnnotate's ability to track temporal information, enabling precise filtering based on annotation dates. This feature also facilitates the comparison of annotations before and after significant cultural moments, such as the COVID-19 pandemic, offering insights into how labels for the same image may evolve over time in response to societal changes.

Scenario 3: Remuneration Schemes in Annotation Situations
For a certain dataset, I want to know which remuneration schemes have been used in annotation situations meant to create annotations for it.
Rationale: This scenario explores the various compensation models employed in annotation situations that have led to annotations for a specific dataset. Identifying remuneration schemes informs us about the motivations and incentives driving annotators, which can impact the quality and consistency of annotations.

Scenario 4: Annotated Entity Types in Annotation Situations
I want to gain insights into the types of entities that have been annotated within

the SituAnnotate ontology. Specifically, I want to know the categories of entities, such as images or documents, that have undergone annotation and are represented in the SituAnnotate KG.
Rationale: This query illuminates the entities whose annotations have been integrated into the SituAnnotate ontology. It offers insight into the categories of entities, such as images and documents, that have undergone annotation and are represented within the SituAnnotate KG. This comprehension is crucial for domain-specific applications as it unveils the breadth of concepts encompassed by the ontology.

Scenario 5: Identifying Annotations based on Lexical Entry
I want to identify all entities that have been annotated using a specific lexical entry, such as "surfboard." Additionally, I want to know the roles that these annotations serve.
Rationale: This question exemplifies how the ontology can be leveraged for the identification of all entities, or entities of a specific type (e.g., images), that have been annotated with the same lexical entry (e.g. "surfboard") and the corresponding annotation roles (e.g., detected object). This query is instrumental in gaining insights into the usage and impact of specific lexical entries across various annotations.

Scenario 6: Identifying Contextual Information for Annotations
For a specific situation in which a lexical entry was used to annotate an entity, I want to know the contextual factors associated with the annotation situation, including the country, date, annotated dataset, remuneration scheme, detection threshold, and details about the annotator.
Rationale: This question aims to provide comprehensive context for a particular annotation scenario, encompassing geographical and temporal aspects, the dataset under annotation, remuneration specifics, detection thresholds, and annotator attributes. It offers a powerful tool for understanding how a ground truth is situated within its originating context.

Scenario 7: Filtering Annotations by Reliability and Roles
I want to filter annotations based on their reliability and roles. Specifically, I want to identify entities with annotations classified under specific annotation roles, such as detected objects or detected emotions, with annotation strengths exceeding certain thresholds. Additionally, I want to know the labels assigned to these entities.
Rationale: This question delves into annotations categorized by specific roles (e.g., detect object, detected emotion, detected action) and their associated annotation strengths. It allows for the filtering of entities based on the reliability or strength of annotations and provides insight into the specific labels.

Scenario 8: Identifying Concepts Typing Annotations about Entities
I want to know the concepts that type annotations for a specific entity. Specifically, I want to know the concepts associated with annotations for the entity, along with their annotation strengths and annotation roles.
Rationale: This scenario focuses on the concepts linked to a particular entity via annotations. It not only provides a list of concepts associated with an entity via situated annotations but also essential details about the nature of these assignments, such as their roles and the strength of these associations. This nuanced view enhances our understanding of the annotations' semantics and reliability.

Scenario 9: Tracking Annotators Responsible for Annotation Labels
I want to identify the annotators responsible for specific labels associated with a particular image. Specifically, I want to attribute annotations to individual annotators, enabling an assessment of their contributions to the annotation process.
Rationale: This scenario delves into the identification of the annotators accountable for specific labels associated with a particular image. This level of detail enables the attribution of annotations to individual annotators, facilitating an assessment of their contributions.

Scenario 10: Artificial Annotators and Shared Model Architectures
I want to explore artificial annotators with shared model architectures within SituAnnotate. Specifically, I want to know what types of annotations about an entity were created by artificial annotators with a specific model architecture. Additionally, for each of these annotators, I want to determine the dataset they were pretrained on, if applicable.
Rationale: This question explores artificial annotators that employ a shared architectural backbone for creating annotations of various types. Identifying shared model architectures sheds light on the integration of automated annotation tools within annotation pipelines. Additionally, it provides insights into the prevalence of specific model architectures and their pretraining on various datasets, contributing to a broader understanding of automated annotation methods.

Scenario 11: Identifying Image Caption Annotations and Annotators
I want to focus on image caption annotations and the annotators responsible for them. Specifically, I want to identify the caption annotations for a specific image and determine who the annotators are for each caption annotation.
Rationale: This query focuses on revealing caption annotations and their respective annotators for a given image. It is vital for examining the generation and attribution of textual descriptions, shedding light on the creators of these annotations and their role in conveying information about the image.

IV.2.2.3 SituAnnotate's Core Concepts

As such, the core goal of the SituAnnotate ontology is to *situate* annotations by connecting them not only to the entity they describe but also to the general situation and to the annotator involved in it. While established ontologies like PROV-O[4] and OpenAnnotation[5] provide robust frameworks for representing provenance and annotations, but they do not specifically address the need to treat the annotation situation as a first-class citizen. Recognizing this gap in existing ontologies, particularly in their ability to facilitate separate queries of entities that are annotation situations distinct from the annotations themselves, SituAnnotate was purposefully designed to fill this gap by introducing three core classes: *Annotation*, *AnnotationSituation*, and *Annotator* (see Figure IV.2.1). This design choice enables SituAnnotate to offer a structured and context-aware representation of annotation situations and their associated entities.

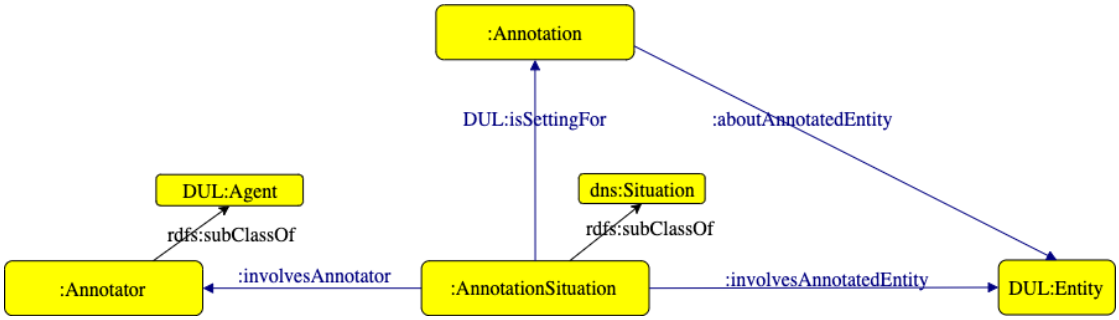


Figure IV.2.1: SituAnnotate at a glance: Core concepts connecting annotations, annotation situations, and annotators.

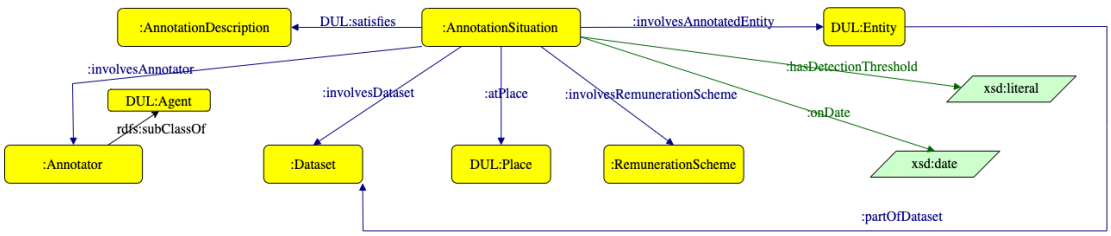


Figure IV.2.2: A detailed view of the SituAnnotate Ontology's core building block, the AnnotationSituation class.

[4] https://www.w3.org/TR/prov-o/. Access date: January 2024.
[5] https://openannotation.io. Access date: January 2024.

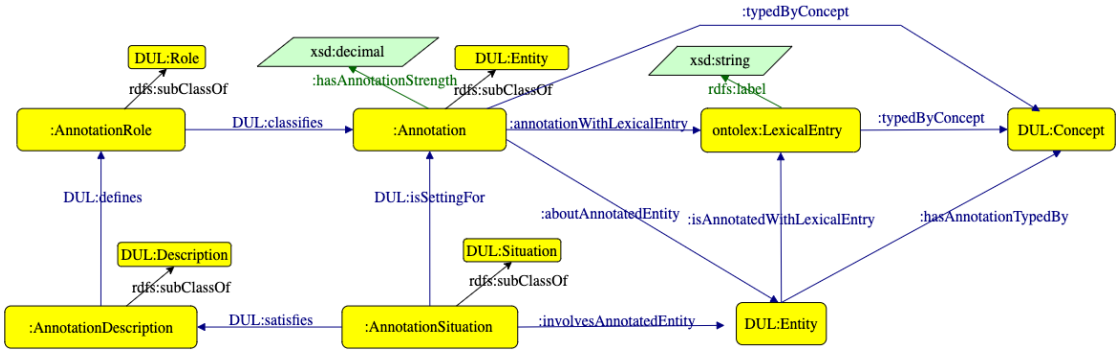


Figure IV.2.3: Deep Dive into the Annotation class: annotation instances connect annotated entities to lexical entries by fulfilling a specific AnnotationRole in a certain AnnotationSituation.

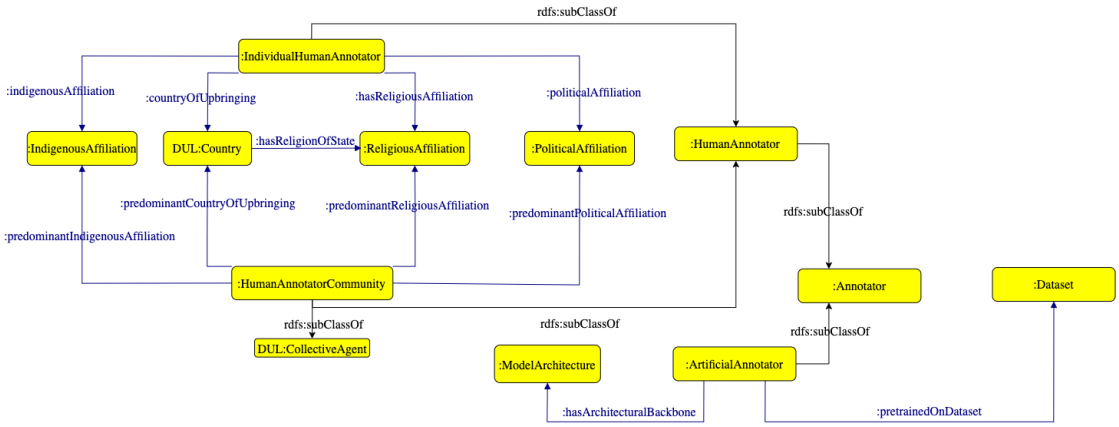


Figure IV.2.4: Deep Dive into the Annotator class: SituAnnotate allows the formal representation of different types of annotators and relevant characteristics that may influence their annotation choices.

IV.2.2.4 Annotation Situations and States of Affairs

The first central Class of the SituAnnotate ontology is the *AnnotationSituation* class, depicted in detail in Figure IV.2.2. This class functions as the cornerstone of the ontology, a subclass of the DUL *Situation* class to maintain alignment with DUL's situational modeling. In essence, the *AnnotationSituation* encapsulates the comprehensive state of affairs in which an annotation may occur: at a precise moment in time, at a specific *Place*, potentially involving a specific *Annotator*, a *RemunerationScheme* by which she is paid for her labor, a *Dataset* to which the *AnnotatedEntity* belongs, and more. Furthermore, each *AnnotationSituation* must adhere to an *AnnotationDescription*. Additionally, this class can incorporate other

pertinent details unique to the situation. By serving as a representation of the contextual environment in which annotations transpire, the *AnnotationSituation* class interconnects all pertinent data, whether contextual or otherwise, associated with the annotation process.

IV.2.2.5 Annotations and Annotation Roles

The second core class is *Annotation*. Instances of this class represent the units responsible for attaching specific meanings, conveyed by lexical units, to an annotated entity in the context of a particular *AnnotationSituation* (see Figure IV.2.3). Annotations are classified by their *AnnotationRole*, a subclass of *Role*. These roles are defined within *AnnotationDescriptions*, adding semantic richness to the ontology, thus enhancing its expressiveness and precision. This approach allows for the representation of diverse annotation types and their roles within the annotation process. Notably, SituAnnotate introduces a distinctive feature where an *Annotation* is a first-order instance capable of establishing relationships with other instances, extending beyond mere textual labels (e.g., "woman," "happiness," or "cemetery"). Instances of the *Annotation* class are not only linked to their corresponding lexical entries but also to the *AnnotatedEntity* they describe (e.g., an image), the specific annotation role they fulfill (e.g., "detected object," "detected emotion," or "detected scene"), the concept typing the lexical entry (e.g., conceptnet:woman), and, importantly, the *AnnotationSituation* within which the annotation originated. This interconnection enables explicit queries to determine the context in which a specific entity was associated with a particular lexical label.

IV.2.2.6 Annotators

The third key class in SituAnnotate is *Annotator*. In this ontology, an Annotator can take one of two forms: an *ArtificialAnnotator*, representing automated programs utilizing a specific *ModelArchitecture* pre-trained on a designated *Dataset*, or a *HumanAnnotator*. The *HumanAnnotator* category is further subdivided into two subclasses: *IndividualHumanAnnotator* and *HumanAnnotatorCommunity*. This differentiation was introduced to handle situations where gathering specific demographic information about individual annotators might be challenging due to privacy considerations. In these cases, data is anonymized by combining and presenting averages. The ontology can accommodate demographic data, such as *PoliticalAffiliation*, *ReligiousAffiliation*, *IndigenousAffiliation*, and country of upbringing. Annotator communities, created by amalgamating data from annotator sets for privacy protection, can also be associated with affiliations using the "predominant" version of affiliation relationships. In essence, this formalization allows for the comprehensive representation of various annotators employed to attribute

meaning to an entity using a lexical label. This flexibility enhances the ontology's capacity to capture the diverse sources and methods used in assigning meaning to entities, including computer vision models, individual annotators, or annotator communities (e.g., the collective annotation provided by the ImageNet dataset annotators).

IV.2.2.7 Aligning with the Dolce Ultra Light Ontology

To ensure the robustness and consistency of the SituAnnotate ontology, it draws inspiration from and aligns with the Dolce Ultra Light (DUL) ontology. By adhering to the principles and design choices of DUL, SituAnnotate benefits from a well-established framework that enhances the ontological modeling of situations, entities, and their relationships. This alignment also promotes interoperability with other ontologies, enabling broader use and integration with existing semantic resources.

IV.2.3 Case Study: Image Annotation Situations

To harness the full potential of our ontology and later assess these scenarios, we expanded our work into a case study focusing on image annotations. These annotations are pivotal in computer vision, a field that stands to benefit significantly from SituAnnotate and our contributions. Computer vision heavily relies on assigning meaning through labels to images, making it particularly susceptible to biases, including human, algorithmic, and interpretational biases [214]. Thus, computer vision serves as an ideal case study, highlighting its heavy reliance on labels and its pronounced vulnerability to concealed biases.

IV.2.3.1 Motivation

In the realm of computer vision, image annotation labels are of paramount importance, serving as the linchpin for understanding, retrieving, and managing the burgeoning volumes of images [230, 311, 378]. These labels, often structured and endowed with semantic meaning through label- and graph-based resources, bridge the chasm between raw image content and its comprehension. Particularly in complex image scenes, the semantic annotation of objects within them empowers automatic understanding and interpretation [309]. Increasingly, linguistic resources and graphs like WordNet [250], ConceptNet [228] and Framester [131] are used to assign and organize labels that give meaning to the raw content of images, for example in the form of scene graphs (e.g., Visual Genome [208]) or taxonomies

(e.g., the Tate collection[6]). These amplify the semantic richness of image features, bolstering image labeling and retrieval systems [304, 314, 352].graph-based models

Critically, the structured representations arising from these annotations also double as invaluable ground truths for the training of computer vision systems, contributing substantially to their precision and efficacy. However, it's imperative to acknowledge that the meanings attributed to images do not exist in a cultural vacuum. Images communicate concepts through a fusion of raw features like lines, colors, shapes, and sizes, alongside culturally coded elements, an aspect that Roland Barthes termed 'connotation' [33]. These coded elements guide human decision-making regarding object identification, labeling, feature ascription, and relationship establishment. In essence, the extraction and portrayal of semantic elements from visual content constitute a code system intricately intertwined with cultural context. This is because *visuality*, different from the purely biological process of vision, is flexible and encompasses "the way that we encounter, look at, and interpret images based on the social, cultural, technological, and economic conditions of their viewing" [146, p. 32]. That is, visuality is a cultural practice with a history marked by different habits or ways of seeing, as well as different types of spectators [124]. This cultural context remains embedded within computer vision pipelines, persisting even in ostensibly straightforward processes like object detection. The 'distant viewing' framework, as introduced by [17], emphasizes the indispensability of a culturally and socially constructed code system to render the semantics of visual content explicit. Labeling and classification systems, though seemingly objective, can inadvertently mirror the values of specific groups or cultures, thereby centralizing power within the process. Despite these intricacies, there lingers a prevailing faith in the objectivity of image labels found in benchmark datasets, often underestimating the cultural and subjective nature of image annotation [146].

IV.2.3.2 The Image Annotation Situation Specialization

We've specialized SituAnnotate to create the Image Annotation Situations (IAS) module, depicted in Figure IV.2.5, with the explicit purpose of tracing the origins of image meanings within culturally coded annotation contexts and facilitating their comparison. This approach is rooted in the notion that an image's semantic labels depend on the specific annotation situation under which it is interpreted. In the IAS module, image annotation is recognized as a contextual situation, similarly to [361], represented by the class `ImageAnnotationSituation`. This context encapsulates all entities relevant to the annotation process, including the image, an-

[6] https://github.com/tategallery/collection/issues/27. Access date: December 2023.

notator, annotation time, location, remuneration details, dataset creation purpose, and more. By applying the Situation pattern, the `ImageAnnotationSituation` class provides a structured framework for contextualizing these entities, allowing for shared features such as location, time, view, causality, and systemic dependencies to be captured.

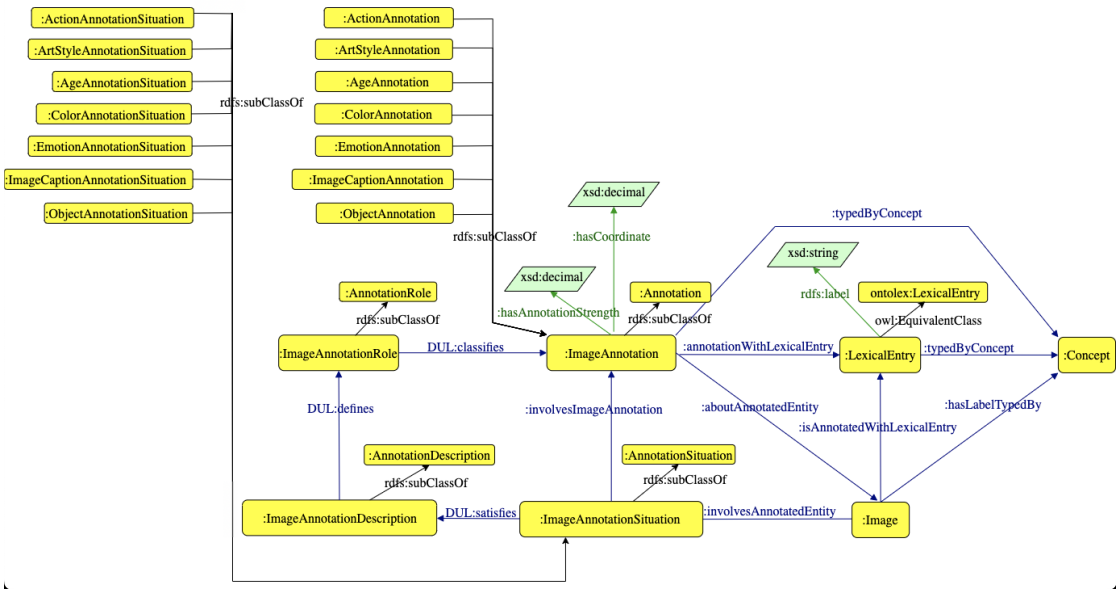


Figure IV.2.5: Specialization of the SituAnnotate pattern specifically for Image Annotation Situations (IAS), crucial in the field of Computer Vision (CV). Further modular specializations can be applied to capture details specific to certain types of annotation situations, such as object detection.

The IAS module emphasizes the need to describe an annotation situation through an `ImageAnnotationDescription`. This description defines the roles and concepts that participate in the state of affairs. The IAS module not only incorporates cultural contextual information regarding image annotation situations but also facilitates comparison between different annotation situations associated with the same image object. This enables users to query and analyze the contexts in which potentially contradictory interpretations of the same image were produced.

Furthermore, SituAnnotate's IAS module already includes classes to support various types of annotations and annotation situations, such as art style annotations, color annotations, object annotations, action annotations, emotion annotations, caption detection, and more. Furthermore, the ontology accommodates detailed annotations, including the assignment of labels to specific regions within

an image using the property `:hasCoordinate`. This feature enables the representation of bounding box annotations for pixels within an image.

Additionally, SituAnnotate offers different image annotation descriptions for the mentioned annotation types (e.g., emotion, color, object, action, caption). These descriptions provide a structured framework for incorporating new annotations into the KG as long as they adhere to the specified description criteria. This flexibility ensures that the ontology remains adaptable and capable of accommodating diverse image annotation data.

IV.2.4 Evaluation

The evaluation protocol consists of several steps aimed at assessing the performance and capabilities of the SituAnnotate system. These steps include the formulation of specialized competency questions, the creation of a toy dataset, the translation of the CQ questions into SPARQL queries, and the execution of these queries over the toy dataset.

IV.2.4.1 Competency Questions (CQs) SPARQL Queries

In the context of the user requirement scenarios, we formulated a set of specialized Competency Questions (CQs). These CQs were designed to reflect the real-world information needs arising from the specific case study and scenarios presented earlier. These questions serve as a valuable tool for assessing the capabilities and performance of the SituAnnotate system in addressing practical use cases. Below, we present the list of CQs derived from our case study and scenarios:

1. CQ1: Which countries have been the location of annotation situations, how many annotation situations were located in each country, and which country has been the location for the highest number of annotation situations?

2. CQ2: Between the years 2020 and 2024, in which annotation situations has the image with ID “ARTstract_14978” been involved?

3. CQ3: What remuneration schemes have been used in annotation situations involving the “ARTstract” dataset?

4. CQ4: What types of entities have been annotated?

5. CQ5: Which images have been annotated using the lexical entry “surfboard,” and what role did these annotations serve?

6. CQ6: For the specific situation in which "surfboard" was used to annotate the image with ID "ARTstract_14978," what contextual factors were associated with the annotation situation?

7. CQ7: Which images have annotations classified under the role of "detected emotion" with an annotation strength exceeding 0.85, and what labels have been assigned to them?

8. CQ8: What concepts type annotations about the image with ID "ARTstract_14978"?

9. CQ9: For each lexical entry (label) that the image with ID "ARTstract_14978" was annotated with, who was the Annotator that assigned that label?

10. CQ10: What types of annotations about the image with ID "ARTstract_14978" were all done by artificial annotators with the "visual transformer" model architecture?

11. CQ11: What are the caption annotations for the image with ID "ARTstract_14978," and who are the annotators responsible for each caption annotation?

IV.2.4.2 Toy Dataset Creation

To evaluate the capabilities and performance of the SituAnnotate system, we crafted a toy dataset in the form of a KG. This dataset emulates real-world scenarios involving multiple annotation situations for a single image, offering a comprehensive testbed for our system. The toy dataset encompasses a diverse array of annotation types, such as object detection, actions, emotions, art styles, colors, and more, all meticulously generated by distinct artificial annotators. To formalize the dataset, we employed the SituAnnotate ontology, ensuring the preservation of extensive information pertaining to each annotation situation. This encompassed details like geographical location, temporal specifics, annotated datasets, remuneration structures, detection criteria, and detailed annotator profiles. This rich contextual data not only enhances the semantic content of the dataset but also enables structured representation for diverse analytical purposes.

Image Data KG An RDF file,[7] it provides comprehensive data for a set of annotations related to a specific image (*:ARTstract_14978*). Each annotation within the dataset is associated with the annotation situation in which it took place.

[7] https://github.com/delfimpandiani/situAnnotate/blob/main/tests/toy_dataset/images_kg.ttl. Access date: December 2023.

These annotations span various dimensions, including actions, age groups, artistic styles, colors, emotions, human presence, image captions, and objects, all linked to relevant ConceptNet concepts. Moreover, each annotation is enriched with an annotation strength value, reflecting its confidence.

Annotation Situations KG This RDF file[8] contains detailed representations of the annotation situations, including details about geographical locations, dates, annotators, and more. Notably, this KG incorporates further information about the artificial annotators used for generating annotations. These annotators are associated with specific model architectures and datasets.

IV.2.4.3 Translation into and Execution of SPARQL Queries

These CQs were subsequently translated into SPARQL queries, creating a formal means to retrieve specific information from the toy dataset. To evaluate the performance and effectiveness of the SituAnnotate system, we executed these SPARQL queries over the toy dataset. For executing the SPARQL queries, we used Ontotext GraphDB,[9] a highly efficient and robust graph database with RDF and SPARQL support. We ran GraphDB in a Docker container, as provided on Github.[10] This platform facilitated the execution of SPARQL queries and retrieval of structured data in accordance with the SituAnnotate ontology. More details are available in Section V.1.5.3 of the Appendix, which summarizes the Competency Questions (CQs) along with their corresponding SPARQL queries and whether they were successfully executed ("Pass" status) in evaluating the SituAnnotate system's performance.

Results

All 11 competency question verification tests were successfully passed, with the expected outcomes matching the actual results. Comprehensive details regarding the results can be accessed in our SituAnnotate GitHub repository[11] and in Section V.1.5.3 in the Appendix. The repository provides in-depth insights into the query outcomes, presenting the retrieved information relevant to each specialized

[8] https://github.com/delfimpandiani/situAnnotate/blob/main/tests/toy_dataset/situations_kg.ttl. Access date: December 2023.

[9] https://graphdb.ontotext.com/documentation/10.0/index.html. Access date: December 2023.

[10] https://github.com/Ontotext-AD/graphdb-docker. Access date: December 2023.

[11] https://github.com/delfimpandiani/situAnnotate/blob/main/tests/competency_question_verification/Results.md. Access date: December 2023.

competency question. We invite readers to visit our online repository for thorough access to our evaluation process and results, facilitating further analysis as needed.

IV.2.4.4 Human-Readable Explanations

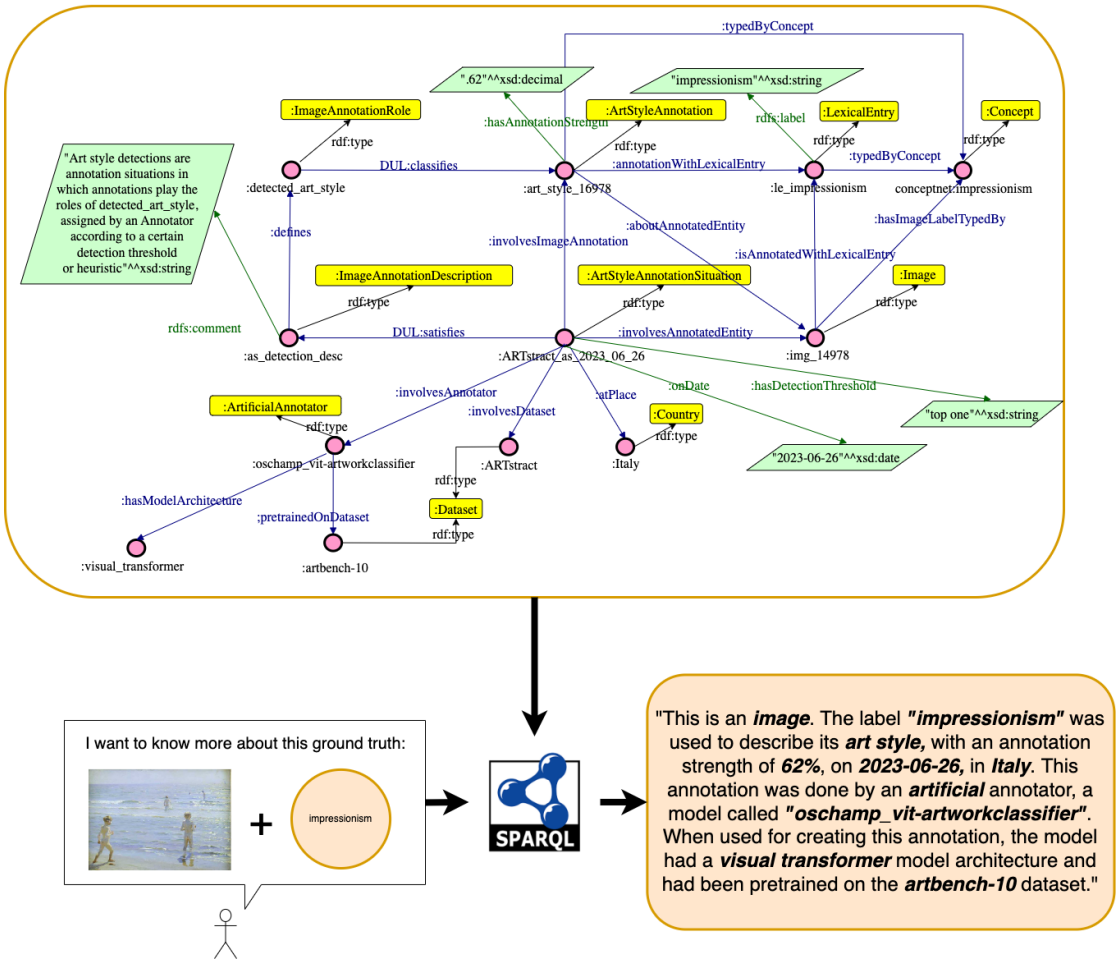


Figure IV.2.6: Illustration of an exemplary usage of the ImageAnnotationSituation specialization, depicting the formal representation of the contextual situation in which the label "impressionism" was assigned to a specific image. Users can employ a single SPARQL query and a small Python script to obtain fully human-understandable statements situating the ground truth label assigned to an entity, as demonstrated in this case with an image.

In addition to assessing the system's performance against the CQs, we aimed to harness the ontology's potential for retrieving situational knowledge regarding

specific ground truths and presenting it in a readily understandable human language. To achieve this objective, we crafted a specialized SPARQL query capable of retrieving comprehensive contextual information for a given entity and label of interest. Subsequently, we developed a Python script to automatically translate the SPARQL query output into a human-readable narrative that elucidates the context surrounding the annotation. This endeavor underscores our commitment to facilitating seamless comprehension and transparency in annotation contexts. The Python script for executing and translating the SPARQL query into natural language is available online.[12]

IV.2.4.5 Results

To demonstrate the practical application of SituAnnotate's capabilities, we selected an image from the toy dataset along with a random label. We executed the aforementioned SPARQL query and a Python script, which translated the query output into a human-readable explanation of the context of the annotation. Figure IV.2.6 provides an example of such an explanation for an image labeled as "impressionism." The figure showcases how the SituAnnotate ontology clarifies and enriches image annotations. It details the context of the "impressionism" label assignment, including information about the annotator, model architecture, and dataset used for training. This practical demonstration highlights the ontology's ability to provide insightful and human-understandable explanations of annotations, making it a valuable tool for situating annotations within their context.

IV.2.5 Discussion

The positive evaluation results highlight the robustness and power of SituAnnotate in formally representing information. These results indicate that SituAnnotate excels in several key aspects, providing significant advantages over traditional annotation methods.

Contextual Precision SituAnnotate provides a highly nuanced and context-aware representation of annotations. By connecting annotations not only to the described entity but also to the broader situational context and the annotator involved, it enables a richer understanding of the circumstances in which annotations are made. This contextual precision is often lacking in traditional annotation approaches that focus solely on labels or strings.

[12] github.com/delfimpandiani/situAnnotate/tests/. Access date: December 2023.

Semantic Enrichment and Expressiveness Unlike traditional annotation methods that often involve fixed annotation tasks and roles, SituAnnotate offers the flexibility of associating the same entity with multiple labels through various "AnnotationRoles." This semantic depth significantly enhances the ontology's expressiveness and precision. Annotators can provide richer and more detailed information about the same entity, enabling a more comprehensive understanding of the annotated data. This flexibility is particularly valuable when dealing with complex or multifaceted annotations. SituAnnotate can even formally represent cases in which the same entity, e.g. an image, is associated with the same label, e.g. "happiness," through various annotations fulfilling different AnnotationRoles, such as detected emotion or detected abstract concept. This semantic depth significantly enhances the ontology's expressiveness and precision

Flexibility in Annotator Representation SituAnnotate is adaptable to various annotator types, including both artificial and human annotators. This adaptability addresses privacy concerns by allowing the aggregation of demographic data when needed. In today's diverse annotation landscape, which includes computer vision models, individual human annotators, and annotator communities, SituAnnotate ensures that all these entities can be formally represented. This reflects the multifaceted nature of modern annotation processes and supports inclusive and diverse annotation practices.

Automated Reasoning and Data-Driven Decision-Making SituAnnotate's structured representation of annotation situations facilitates automated reasoning through SPARQL queries and semantic technologies. Machines can infer relationships, make connections, and retrieve information efficiently, streamlining the annotation understanding process. This automation not only saves time but also supports data-driven decision-making. Users can leverage SituAnnotate to extract valuable insights and patterns from annotated data, enabling evidence-based decisions and enhancing the utility of labeled datasets. Moreover, SituAnnotate enables reasoning over various aspects of annotated data. For example, it allows for reasoning over whether images tagged by models with certain architectures display the co-occurrence of certain objects, or whether certain macro-categories (e.g., animals) are present. This can be achieved through reasoning over the connected resources, like WordNet, offering a deeper understanding of the annotated content and potentially revealing hidden relationships within the data.

Enhanced and User-Friendly Human Understanding SituAnnotate, despite its machine-readable foundation, offers a user-friendly ontological framework that remains accessible to researchers, domain experts, and annotators alike. This

approach ensures that the ontology isn't confined to machines but serves as a valuable resource for human understanding. Moreover, the integration of SPARQL queries and Python scripts empowers users to effortlessly access and interpret situational knowledge tied to specific annotations. This user-friendly feature enhances transparency and facilitates comprehension, making SituAnnotate a versatile resource catering to both machine-driven AI technologies and human expertise. This symbiotic relationship fosters a deeper synergy between AI capabilities and human insights, emphasizing the ontology's significance in bridging the gap between technology and human cognition.

Comparing Annotation Situations for Enhanced Understanding SituAnnotate's ability to represent multiple labels and annotation situations related to the same annotated object provides users with a powerful tool for enhanced understanding. Through SPARQL queries, users can retrieve all AnnotationSituations for an object, enabling detailed comparisons of potentially conflicting interpretations. This feature enhances the understanding of diverse annotation contexts and their implications, supporting better decision-making and data analysis.

Mitigating Bias and Enhancing Ethical AI SituAnnotate serves as a robust tool in the battle against bias through its context-aware data annotation capabilities. Annotators can furnish essential details about data sources, annotator demographics, and the rationale behind labeling decisions. This wealth of contextual information empowers AI developers to scrutinize and rectify any latent biases when examining annotated data. By doing so, SituAnnotate champions transparency and fairness throughout the data annotation process. It contributes significantly to the ethical dimensions of AI development and deployment. As ethical considerations in AI data labeling take center stage, SituAnnotate stands as a pivotal asset, providing a context-aware framework for the meticulous recording and management of annotations.

IV.2.5.1 Limitations

Despite its promising capabilities, SituAnnotate does have some limitations and challenges:

1. **Knowledge Representation Overhead:** While SituAnnotate offers enhanced contextual knowledge representation, this also introduces an overhead in terms of ontology creation, maintenance, and population. It may require substantial time and effort to initially set up and continuously update.

2. **Capturing Human Subjectivity and Cultural Nuances:** One notable challenge lies in the complexity of capturing the full scope of contextual factors that affect human subjectivity and the diverse cultural nuances that can influence annotations. While SituAnnotate offers a structured framework, it does not fully capture the richness of human interpretation.

3. **Scalability Concerns:** SituAnnotate's scalability may be a concern when applied to massive datasets, where managing a vast number of annotation situations and annotators can become unwieldy. Optimizing the ontology for large-scale applications is an ongoing challenge. This is also because the use of SPARQL queries and scripts to retrieve human-understandable explanations can be resource-intensive.

4. **Privacy Mechanisms:** SituAnnotate's ability to address privacy concerns may require further refinement to provide more robust mechanisms for data anonymization and aggregation. Ensuring the privacy and confidentiality of sensitive data is crucial.

These limitations should be considered when implementing SituAnnotate in real-world scenarios, and ongoing research and development efforts may help mitigate some of these challenges.

IV.2.5.2 Further Directions

As SituAnnotate continues to evolve, there are several avenues for further research and development:

- **Usability Improvements:** Prioritize creating user-friendly tools and interfaces that simplify the process of integrating SituAnnotate into annotation workflows. Consider developing user-friendly graphical user interfaces (GUIs) for creating and querying annotations, enhancing accessibility for a broader user base.

- **Scalability:** Investigate methods to enhance SituAnnotate's scalability, particularly when dealing with large datasets. This may involve optimizing SPARQL queries or exploring distributed computing solutions to handle increasing volumes of data efficiently.

- **Enhanced Automation:** Continue to advance automation tools for generating human-readable explanations from the ontology. Explore Natural Language Processing (NLP) techniques to produce more coherent and concise explanations, reducing the need for manual intervention.

- **Interoperability:** Ensure that SituAnnotate remains compatible with other ontologies and standards in the data annotation and semantic web domain. Seamless integration with existing systems is essential for broader adoption.
- **Community Involvement:** Foster collaboration and engagement within the research community to refine and expand SituAnnotate. An active user and developer community can drive further innovation and adoption. Additionally, seek collaboration with the global research community to address cultural biases and diversify the ontology's applicability.
- **Ethical Considerations:** Delve into the ethical implications of SituAnnotate's real-world applications, particularly concerning privacy, bias, and transparency. Develop comprehensive guidelines and best practices for ethical annotation processes, promoting responsible AI development.
- **Scenarios and Use Cases:** Continue to develop and document a diverse set of real-world scenarios and use cases where SituAnnotate has demonstrated its practical value. Providing concrete examples can help potential users grasp its applicability better.
- **Integration with AI Systems:** Explore seamless integration possibilities of SituAnnotate with AI systems, particularly in domains like computer vision, natural language processing, and KGs. Incorporating advanced techniques for handling multi-modal data, including text, images, and videos, can broaden its applicability.
- **AI Ethics and Fairness:** Investigate how SituAnnotate can be integrated with emerging AI ethics and fairness frameworks. Contributing to more responsible and equitable AI development aligns with the growing importance of ethical considerations in the field.

IV.2.6 Conclusions

In conclusion, the SituAnnotate ontology provides a robust and context-aware framework for situating ground truths, i.e., representing annotations within the contextual situations from which they arise. Aligned with the Dolce Ultra Light ontology, it ensures consistency and interoperability, while its expressive relationships and semantic depth enhance annotation context understanding. Researchers and practitioners can use SituAnnotate to model, analyze, and interpret annotations in a structured and standardized way, making it a valuable contribution to data annotation and knowledge representation. SituAnnotate overcomes traditional annotation method limitations, benefiting both human annotators and

automated processes with a structured, machine-readable format that remains human-readable. Its SPARQL query support enables efficient data retrieval and analysis, bridging the gap between structured data and human comprehension, enhancing annotation efficiency and accuracy, and promoting transparency and ethical considerations in data annotation—a crucial step for responsible AI development. Ultimately, SituAnnotate's contextual annotations enhance AI decision-making, aiding models in adapting to real-world scenarios and advancing ethical AI implementation.

Chapter IV.3

Stitching the Gaps with Situated Perceptual Knowledge

Summary This chapter presents a novel approach to AC image classification, leveraging *situated perceptual knowledge* through the ARTstract Knowledge Graph and Knowledge Graph Embeddings (KGE). Our aims are two-fold: establish a reasoning-enabling knowledge graph to deepen our grasp of AC evocation in ARTstract, and employ this knowledge to enhance AC image classification performance and explainability. This chapter outlines the development of the AKG, which integrates perceptual semantics and SituAnnotated metadata, linking images to perceptual concepts while acknowledging subjective bias. Additionally, we enhance the AKG with high-level linguistic frames extracted from image captions. We delve into the transformation of the AKG into KGE for the purpose of conducting AC image classification experiments. These experiments encompass both KGE-only and hybrid approaches, aiming to examine different fusion strategies for combining KGE with Vision Transformers (ViT). Some of these approaches involve the utilization of relative representations [261] to integrate the two embedding types while preserving invariance in the face of latent space transformations. Our interpretability results reveal that ViT excels in low-level visual attributes like colors and textures, while KGE demonstrates proficiency in capturing more abstract and semantic scene elements, highlighting the contrasting capabilities of these two embedding methods in deciphering high-level semantic elements. Critically, our hybrid methods outperform existing state-of-the-art techniques, with the synergy between the situated perceptual knowledge of the KGE embeddings and the sensory-perceptual understanding acquired by ViT leading to superior performance compared to deep and classical machine learning approaches. These results collectively underscore the potential of neuro-symbolic methods in providing robust image representation for intricate visual comprehension tasks.

IV.3.1 Introduction and Background

In the rapidly evolving field of CV, the enduring challenge is to equip machines with human-like cognitive capabilities, surpassing data-driven pattern recognition to bridge the gap between bottom-up signal processing and top-down knowledge retrieval and reasoning [3]. This goal is rooted in the understanding that "humans are not mere appearance-based classifiers; we acquire knowledge from experience and language" [239]. While explicit knowledge has historically been recognized as a way to improve automatic image understanding, modern data-driven techniques aim to acquire the majority of this knowledge from the training data itself.

Nonetheless, situations arise where annotated data is scarce or simply not enough, leading to the demand of methods to incorporate both spatial and semantic reasoning for advancing the next generation of vision systems [78]. A key potential solution is found in the convergence of symbolic, knowledge-driven AI that prioritizes knowledge representation and sub-symbolic, data-driven AI rooted in machine learning. The synergy between these paradigms holds the potential to yield more intelligent, hybrid systems [38]. Consequently, the development within CV of methods for leveraging textual background knowledge [4] and integrating reasoning has gained substantial attention. These efforts often involve KGs: structured databases that represent entities and their relationships in directed, edge-labeled graphs, often adhering to an ontological schema and semantic web standards like Linked Data[1] [357].

In the context of replicating human-like vision for complex tasks such as AC image classification, it is essential to mimic human perceptual knowledge when relying on visual information. This involves recognizing perceptual semantics, including objects and colors (as discussed in Chapter III.2). Additionally, comprehending symbolic representations within images depends on grasping common-sense associations [191]: an individual can recognize that a "cat" falls into the more abstract category of "animal," or that a "car" is associated with the frame of "transportation." Our results in Chapter IV.1, showed that we can automatically establish these connections with high-level linguistic frames by reifying perceptual semantic descriptors. Furthermore, individuals possess embodied knowledge that influences contextual meaning attribution. Chapter IV.2 illustrated encoding this bias-awareness into a machine-readable format, combining semantics with contextual factors called 'annotation situations.' For example, different artificial detectors introduce biases akin to an 'embodiment' shaped by architecture and pretraining. Annotating their architecture and pretraining data can equate machine vision's embodiment with human vision, enabling direct parallels. This contextual knowledge enhances the alignment with human vision and cognition.

[1] https://www.w3.org/wiki/LinkedData. Access date: December 2023.

Based on these insights, this chapter integrates knowledge from Chapters II.2, III.2, IV.1, and IV.2 into our *situated perceptual knowledge* paradigm. We build the ARTstract-KG, a resource enabling automatic reasoning, situating, and linking of ARTstract images with perceptual semantics, commonsense knowledge, and ACs in a KG, leveraging the *SituAnnotate* ontology introduced in Chapter IV.2. By embedding the KG into a vector space, we create image vectors informed by the situated knowledge graph data. These representations are subsequently employed in the AC image classification task, for which we investigate hybrid methods to bridge the gap between the end-to-end deep vision approaches and the situated perceptual knowledge paradigm. In summary, this chapter accomplishes the following:

- **Introduction of ARTstract-KG**: We introduce the AKG as a resource that combines perceptual semantics from ARTstract images with cultural and commonsense symbolic knowledge.
- **KG Embedding Representations**: We create image representations by embedding ARTstract-KG into a vector space. These representations incorporate situated knowledge graph information and are utilized in the AC image classification task.
- **Hybrid Fusion**: We present hybrid methods aimed at bridging the gap between the end-to-end deep vision paradigm and the situated perceptual knowledge paradigm. Our exploration includes the fusion of both absolute and relative representations [261].
- **Interpretability Experiments**: To comprehend the inner workings of our embedding models, we probe relevant similarities with training instances and qualitatively analyze them to grasp the models' abilities to capture symbolic and embodied aspects of image content.

In Section IV.3.2, we introduce our idea of situating and formalizing extracted perceptual semantics of images into a KG and using derived embeddings from it for enhancing AC image classification. In Section IV.3.3, we delve into the construction of the ARTstract-KG (AKG) via the use of the SituAnnotate ontology to seamlessly integrate extracted perceptual semantics into a knowledge-driven resource (see Subsection IV.3.3.1. That section also describes the process of embedding the AKG (see Subsection IV.3.3.2), the utilization of these embeddings for classification purposes (see Subsection IV.3.3.3), and our interpretability experiments (see Subsection IV.3.3.5). Section IV.3.4 presents the resulting ARTstract-KG, as well as the performance and explanability results for all the tested AC image classification methods. In Section IV.3.5, we analyze and discuss the achievements of this chapter.

IV.3.1.1 Background: Knowledge-Based Reasoning in CV

Background knowledge can be integrated into computer vision pipelines in various ways: preprocessing knowledge to augment input, incorporating knowledge as embeddings, post-processing through explicit reasoning mechanisms, and using knowledge graphs to influence neural network architectures [4]. Knowledge and reasoning have been used in computer vision for decades now. Markov Logic Networks (MLN) [295], which uses weighted First Order Logical formulas to encode an undirected grounded probabilistic graphical model, used by [408] to reason about object affordances (ex., *fruit is edible* or *basketball is rollable and round*). Probabilistic Soft Logic also uses a set of weighted First Order Logical rules, used to declare a Markov Random Field, and has been used by [231] to detect collective activities such as *crossing, queuing, waiting,* and *dancing* in videos. Description Logics [23] models relationships between entities in a particular domain, and has been used to reason and check consistency on object-level and scene-level classification systems, such as in [97].

Many efforts in the utilization of knowledge bases for image classification have concentrated on the fusion of graphs to integrate image-level specifics with general knowledge. KGs capture comprehensive world knowledge, while scene graphs capture the semantic content within an image. In essence, numerous studies have sought to merge visual datasets, like Visual Genome [209], with extensive knowledge bases that offer both commonsense information pertaining to visual concepts and non-visual concepts. For example, ConceptNet [161] is a semi-curated multilingual Knowledge Graph that encodes commonsense knowledge about the world and was built primarily to assist systems that attempt to understand natural language text. The nodes of the graph are concepts–words or short phrases written in natural language. Edges are labeled with meaningful relations, such as <reptile, isA, animal>, and each edge has an associated confidence score. ConceptNet is semi-curated, so it has large coverage but less noise than other resources. It has been exploited in computer vision, including by [216], who use the commonsense knowledge encoded in ConceptNet to enhance a language model and apply this knowledge to two recognition scenarios (action recognition and object prediction).

ConceptNet has also been exploited by the work most related to ours: the authors in [191] propose a neural network framework named SKG-Sym (scene and knowledge graph symbolic image detection), an approach that integrates general knowledge and the visual components of scene graphs in the learning process. They use graph convolutional networks over each graph to assign weights to visual components and knowledge concepts. They also employ a scene graph detector to integrate. For each image, they create a knowledge graph that extracts general knowledge of each detected object, keeping ConceptNet edges such as relatedTo, isA, partof, madeof, atlocation, etc [191].

With the introduction of the Gated Graph Neural Network (GGNN) [224], specifically designed for graph-structured data, and utilizing message passing with gating mechanisms to update node representations based on information from neighboring nodes, novel approaches for integrating graph knowledge into visual reasoning have emerged. Marino et al. [239] introduce the Graph Search Neural Network (GSNN) as an efficient means to incorporate extensive knowledge graphs into a vision classification pipeline. Their network learns a propagation model that can reason about various types of relationships and concepts, generating node outputs used for image classification. Notably, this approach addresses computational challenges associated with GGNNs for large graphs, enabling efficient training for image tasks that leverage extensive knowledge graphs. An additional key feature is its capability to provide explanations for classifications by tracing the propagation of information within the graph. The authors in [156] introduced a Graph Neural Network (GNN) for image understanding, which surpasses traditional feature and decision fusion approaches by recognizing the potential for features to interact and exchange information. Their model was applied to two image understanding tasks, specifically group-level emotion recognition (GER) and event recognition, both of which demand semantic sophistication and the interaction of multiple deep models for the synthesis of various cues. Notably, this approach achieved state-of-the-art performance in these image understanding tasks.

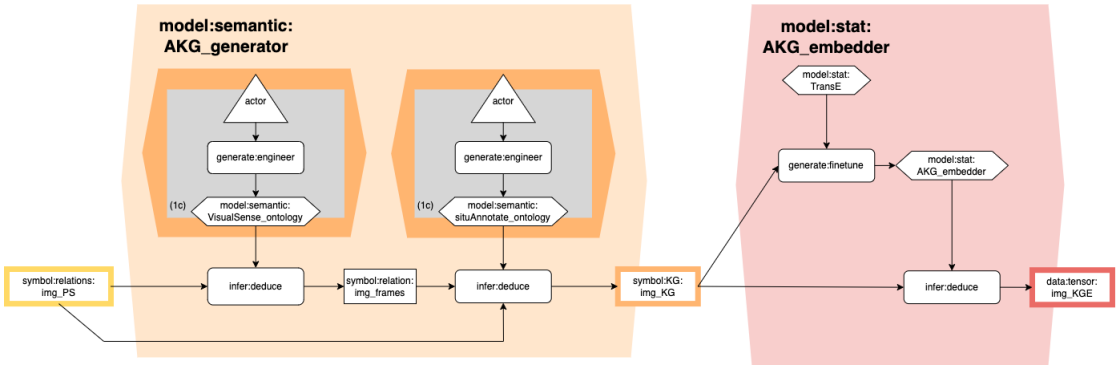


Figure IV.3.1: We create graph representations for images by analyzing their perceptual details, annotations, and high-level linguistic frames. Each image is represented as a node in the ARTstract-KG.

IV.3.2 Idea: Situated Perceptual Knowledge (SPK)

Our overarching goal is to enhance the process of AC image classification by incorporating situated perceptual knowledge. This relies on two main components of this chapter: the creation of the AKG to represent it, and then its embedding to create a novel vector image representation used for the downstream task of AC image classification.

IV.3.2.1 ARTstract KG Creation and Embedding

This method involves the transformation f_{KG} of an image's perceptual semantics representation, (I_{PS}) to a Graph representation incorporation situational metadata and commonsense knowledge. The AKG is created via the SituAnnotate framework, allowing images to be represented as nodes of the AKG, denoted I_{KG}. Then, we embed the KG, resulting in image representations that are KGE, denoted I_{KGE}.

$$f_{KG} : I_{PS} \rightarrow I_{KG} \subseteq \mathbb{G} \tag{IV.3.1}$$

$$f_{KGE} : I_{KG} \rightarrow I_{KGE} \subseteq \mathbb{R}^{128} \tag{IV.3.2}$$

IV.3.2.2 AC Image Classification Using the KG Embeddings

A crucial point in this chapter is to show how these KGE image representations (I_{KGE}) can be used in multiple ways for the task of AC image classification. This chapter addresses two overarching paradigms: the Situated Perceptual Knowledge approach (SPK), which only relies on the KGE representation, and the Hybrid paradigm which exploits both the (I_{KGE}) representation and the deep features I_{DL} from Chapter II.2.

1. KGE-Only Image Classification

We employ the ARTStract-KG derived image representations (I_{KGE}) in a classification model referred to as the "KGE-Only" method (see Figure IV.3.2). This approach exclusively utilizes KGE embeddings for images to train a Multi-Layer Perceptron (MLP) classification model:

$$\hat{y} = \arg\max(p(y_i | I_{\text{KGE}}, \theta)) \tag{IV.3.3}$$

2. Hybrid Approach with KGE and ViT

In our efforts to merge the strengths of the paradigms we have investigated so far, we explore two hybrid approaches that involve combining I_{KGE} representations with I_{DL} features into a new, hybrid image representation I_{H} (see Figure IV.3.3).

Simple Concatenation The first approach involves a straightforward concatenation of the two embeddings to represent an image.

$$f_H : [I_{KGE} \subseteq \mathbb{R}^{128}; I_{DL} \subseteq \mathbb{R}^{768}] \rightarrow I_H \subseteq \mathbb{R}^{896} \tag{IV.3.4}$$

Relative Representations Our approach involves transforming the two vector image representations, I_{KGE} and I_{DL}, into their relative versions, $I_{\text{R-KGE}}$ and $I_{\text{R-DL}}$ and exploring hybrid methods to combine the relative embeddings into a hybrid image representation I_{H}.

$$f_{R-KGE} : I_{KGE} \subseteq \mathbb{R}^{128} \rightarrow I_{R-KGE} \subseteq \mathbb{R}^{|A|} \tag{IV.3.5}$$

$$f_{R-DL} : I_{DL} \subseteq \mathbb{R}^{768} \rightarrow I_{R-DL} \subseteq \mathbb{R}^{|A|} \tag{IV.3.6}$$

We adopt relative representations [261] to represent each training sample with respect to a set of anchors. A subset A of the training data X is selected as anchor samples, each training sample is represented with respect to the embedded anchors $e_{a^{(j)}} = E(a^{(j)})$ with $a^{(j)} \in A$ via a generic similarity function $sim : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$. This yields a scalar score r between two absolute representations $r = sim(e_{x^{(i)}}, e_{x^{(j)}})$. Thus, the relative representation of $x^{(i)} \in X$ is defined as:

$$r_{x^{(i)}} = (sim(e_{x^{(i)}}, e_{a^{(1)}}), sim(e_{x^{(i)}}, e_{a^{(2)}}), \ldots, sim(e_{x^{(i)}}, e_{a^{(|A|)}})) \tag{IV.3.7}$$

These include the concatenation $||$ of the relative embeddings, and element-wise operations $\odot$ on the relative embeddings:

$$f_H : \left[I_{R-DL} \subseteq \mathbb{R}^{|A|}; I_{R-KGE} \subseteq \mathbb{R}^{|A|}\right] \rightarrow I_H \subseteq \mathbb{R}^{(|A|*2)} \tag{IV.3.8}$$

$$f_H : I_{RR-DL} \subseteq \mathbb{R}^{|A|} \odot I_{RR-KGE} \subseteq \mathbb{R}^{|A|} \rightarrow I_H \subseteq \mathbb{R}^{|A|} \tag{IV.3.9}$$

Regardless of the hybrid method chosen to combine I_{KGE} representations with I_{DL}, the resulting I_{H} is employed for classification training and evaluation.

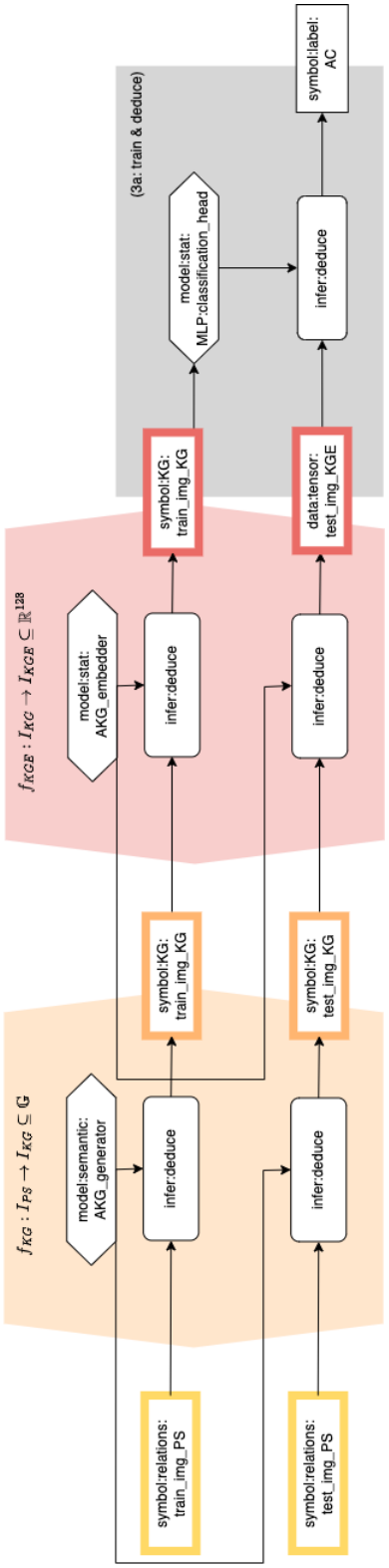


Figure IV.3.2: Architecture of the situated perceptual knowledge approach to AC image classification.

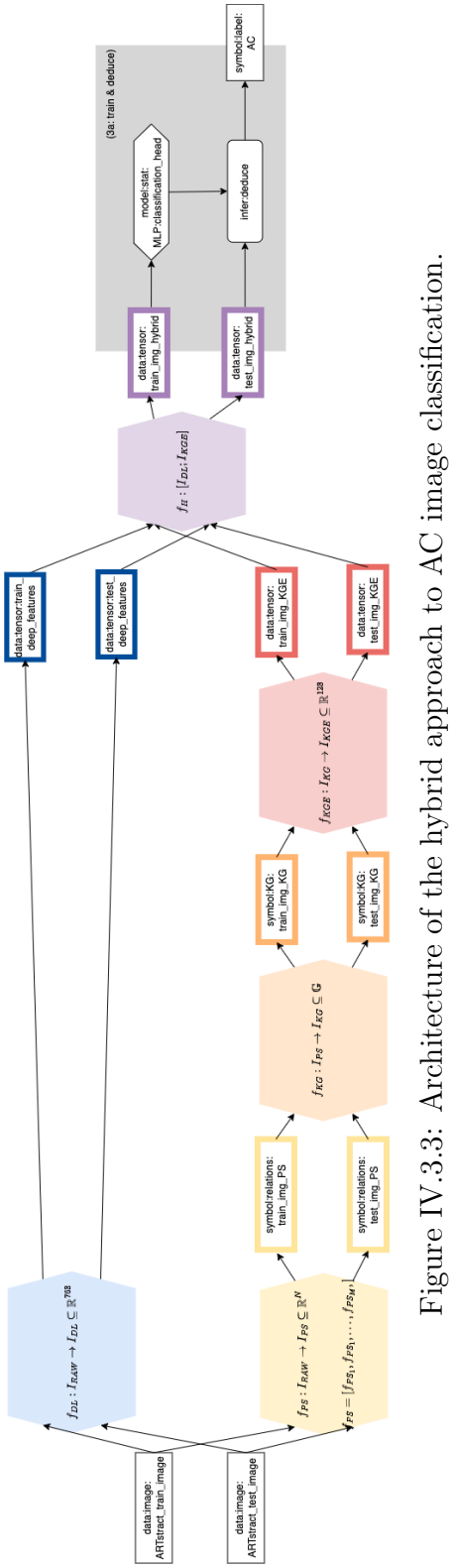


Figure IV.3.3: Architecture of the hybrid approach to AC image classification.

IV.3.3 Approach

IV.3.3.1 ARTstract-KG Construction

In this section, we outline our methodology for constructing the AKG (see Figure IV.3.1), which aims to capture the intricate relationships between perceptual semantics, images, and their contextual annotations. Drawing insights from Chapters III.2 and informed by the findings in Chapter IV.2, we identified that not only the specific perceptual semantic labels (e.g., top colors, actions, objects, etc.) but also the annotation situations in which these labels were assigned to individual images held substantial knowledge value. Additionally, we enrich the KG with commonsense linguistic frames following the results of Chapter IV.1.

Ontology Reuse We harness the *SituAnnotate* ontology, previously introduced in Chapter IV.2. This ontology already has a module specifically tailored for *image* annotation situations, as detailed in the chapter. In line with this specialized module, we opted to embed each chosen label within its respective annotation context. We accessed the ontology directly via its permanent IRI[2]. We also reuse the Framester [131] schema to refer to ConceptNet and WordNet IRIs.

Reification of Annotation Situations and PS Annotations To formally represent the contexts in which perceptual semantic labels were assigned to the 14K+ ARTstract images, we reified them as instances of various subclasses of `ImageAnnotationSituation`. In practical terms, this involved transforming each entry row in Table III.2.1 (as introduced in Chapter III.2) into an instance of `AnnotationSituation`. The resulting triples encapsulate intricate details about these annotation situations, encompassing factors such as geographical locations, timestamps, annotators, specific model architectures, datasets, and more. To provide a tangible example, the following illustrates the triples associated with a single `AnnotationSituation`:

```
:ARTstract_as_2023_06_26 a :ArtStyleAnnotationSituation ;
    :involvesAnnotatedEntity :ARTstract_14978 ;
    :atPlace :Italy ;
    :hasDetectionThreshold "top_one" ;
    :involvesAnnotator :oschamp_vit-artworkclassifier ;
    :involvesDataset :ARTstract ;
    :onDate "2023-06-26"^^xsd:date ;
    :satisfies :as_detection_desc .
```

[2] https://w3id.org/situannotate

```
:as_detection_desc a ns1:ImageAnnotationDescription ;
    rdfs:comment "Art style detections are annotation
        situations in which annotations play the role
        of detected_art_style, assigned by an Annotator
        according to a certain detection threshold or
        heuristic"^^xsd:string ;
    :defines :detected_art_style .

:oschamp_vit-artworkclassifier a :ArtificialAnnotator ;
    :hasModelArchitecture :visual_transformer ;
    :pretrainedOnDataset :artbench-10 .
```

To reify the perceptual semantic labels assigned to ARTstract images, we formally represented them as instances of the *Annotation* class. Triples were constructed to connect each annotation to the corresponding `AnnotationSituation` from which it originated, the associated `Image`, the employed `LexicalEntry`, the assigned `AnnotationStrength`, the classification `AnnotationRole`, and the ConceptNet concept that provided its typification. To exemplify, consider the following set of triples explicitly encoded for a single annotation:

```
:14978_ARTstract_as_2023_06_26 a :ArtStyleAnnotation ;
    :aboutAnnotatedEntity :ARTstract_14978 ;
    :annotationWithLexicalEntry :le_Impressionism ;
    :hasAnnotationStrength 0.6149182915687561 ;
    :isAnnotationInvolvedInSituation :ARTstract_as_2023_06_26 ;
    :isClassifiedBy :detected_art_style ;
    :typedByConcept conceptnet:impressionism .
```

Enrichment with Linguistic Frames To further enhance the AKG, we incorporate high-level linguistic frames extracted from image captions. For each caption, we follow the same extraction procedure as introduced in Chapter IV.1. Using FRED [131], we retrieve WordNet synsets for words in the caption. These synsets are then employed as triggers to collect frames. Both the WordNet synsets and frames are integrated into the knowledge graph (see Figure IV.3.10). This addition contributes to a more comprehensive and expressive representation.

Knowledge Graph Construction The KG was constructed using a Python script[3] that mapped perceptual semantics from a JSON file to the SituAnnotate ontology. RDFlib, a Python library for handling RDF data, was employed. Key steps included:

[3] https://github.com/delfimpandiani/ARTstract-KG/tree/main/ARTstract-KG_creation/ARTstract_kg_construction. Access date: December 2023.

- Creation of triples for Annotation Situations based on information from the situations JSON file.
- Iteratively processing the Perceptual Semantics JSON data to extract details about image annotations, their contexts, strengths, etc.
- Mapping and iterative cleaning for Conceptnet matching.
- Triple definition for relationships between images, annotation situations, annotations, lexical entries, annotation strengths, and ConceptNet concepts.
- Serialization of the RDF graph into Turtle format.

IV.3.3.2 Learning ARTstract-KG Embeddings

Preprocess the KG to TSV Format In the initial phase of KG preparation for training, we transformed the original Turtle (RDF) format KG into a TSV format compatible with the torchkge library. This conversion involved parsing the data into a TSV file, including source entities, relationships, and target entities, using the Pandas library.

Data Leakage Prevention To ensure that the KG remained free from any information that could leak the target AC clusters, we systematically removed any rows including AC cluster names in the subjects or objects. This led to removals of some triples with relations: `:annotationWithLexicalEntry`, `:typedByConcept`, and `:annotationWithEvocationCluster`. Our filtering process was all-encompassing, extending across the entire dataset, spanning all images, and covering each data split, including the train, validation, and test sets. The outcome was a KG that maintained its separation from the target AC clusters, thereby upholding the integrity of the AC cluster ground truth.

Knowledge Graph Embedding To transform the KG in TSV format into KGE, we employed the TransE model, opting for the default hyperparameters to establish a fundamental baseline. This included a fixed random seed of 42, an embedding dimension of 128, and a batch size of $8192 * 4$. During training, we employed the MarginLoss criterion with a margin value of 1 for guidance. Negative samples were generated using a Bernoulli negative sampler with 100 negatives per sample. We trained for 1000 epochs.

IV.3.3.3 AC Image Classification

We conducted several experiments utilizing the KG embeddings (see Table IV.3.1), relying on both original ("absolute") embeddings, and their relative representations [261] (see Figure IV.3.4). Each experiment employed a distinct input embedding to represent the image, but all methods employed a common MLP architecture for the classification head. This architecture consists of two sequential linear layers with a Rectified

Linear Unit (ReLU) activation function and a dropout layer (dropout rate of 0.3) for regularization. The primary variations among these models are in the input vectors used for training, transformation dimensions in the first linear layer, and specific operations applied to entity embeddings. In all architectures, the second linear layer maps the feature representation to the number of output classes. During training, the Cross-Entropy Loss function is employed to compute the error between predicted class scores and actual ground truth labels. A fixed learning rate ($lr = 0.001$) is used, and each architecture is trained for 50 epochs. The efficiency of data processing is enhanced through multi-threading with 16 workers.

Table IV.3.1: Summary of used embeddings and their dimensionality. ||: concatenation; $\odot$: element-wise (Hadamard) product.

Embedding Type	Dimensionality
Absolute KGE	128
Absolute ViT	768
Relative KGE	700
Relative ViT	700
Absolute KGE || Absolute ViT	896
Relative KGE || Relative ViT	1400
Relative KGE || Absolute ViT	1468
Relative KGE $\odot$ Relative ViT	700

Absolute Embeddings

Absolute KGE This model exclusively utilizes the original (called "absolute" KGE learned with TransE (see Figure IV.3.2. The embeddings have a dimensionality of 128, in the MLP they are transformed by the first linear layer into a 64-dimensional space.

Absolute ViT This model exclusively utilizes the ViT embeddings with a dimensionality of 768, just like in Chapter II.2.

Relative Representation Embeddings

We adopt the method of relative representations, introduced by Moschella et al. (2022) [261] to represent each embedding in the training distribution with respect to a set of embedded anchor vectors (see Figure IV.3.4, bottom). A subset A of the training data X is selected as anchor samples, and each training sample is represented with respect to the embedded anchors $e_{a^{(j)}} = E(a^{(j)})$ with $a^{(j)} \in A$. The relationship between the anchors and other samples is captured using a generic similarity function $sim : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, yielding a scalar score r between two absolute representations $r = sim(e_{x^{(i)}}, e_{x^{(j)}})$ (for more details, see Subsection IV.3.2.2).

Relative KGE and Relative ViT These models exclusively utilize the relative versions of the KG embeddings (rel-KGE) or of the ViT embeddings (Rel-ViT), each of which has a dimensionality of 700.

Hybrid Representation Embeddings

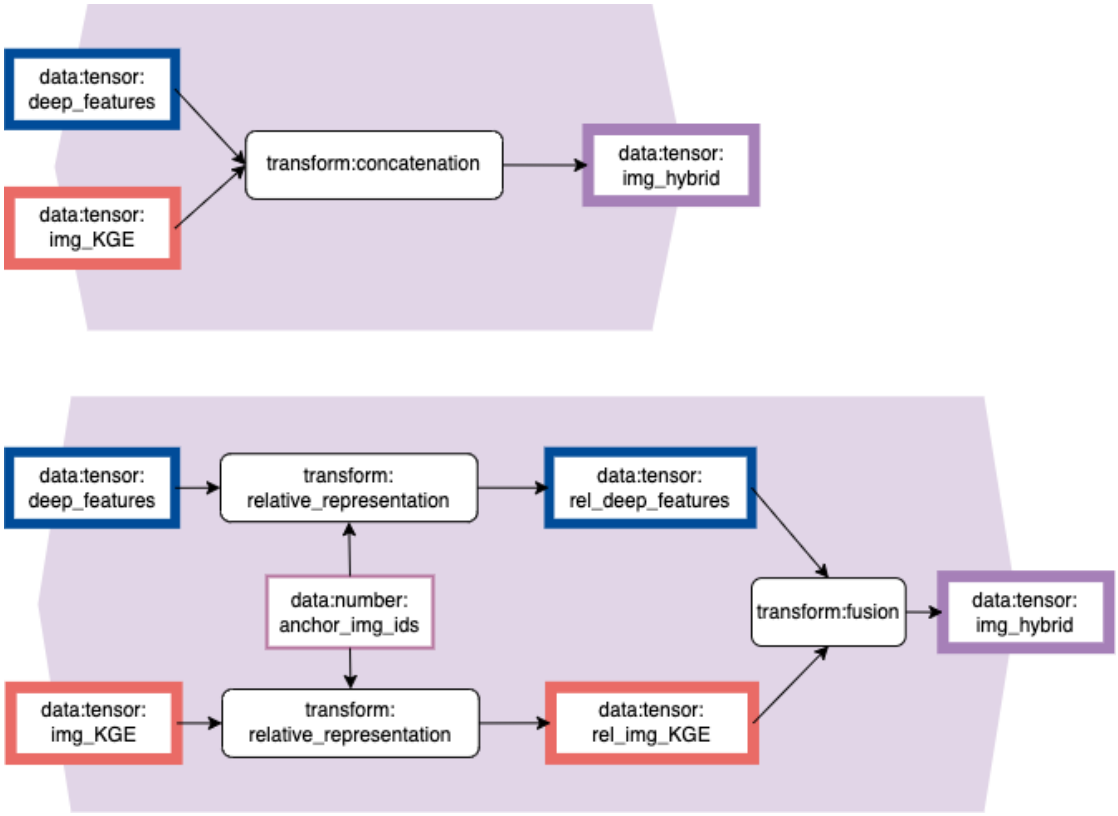


Figure IV.3.4: Approach to fuse deep learning vectors with knowledge graph embeddings. Top: simple concatenation method of absolute embeddings. Bottom: relative representations [261] are first calculated with respect to a set of anchors, and then they are fused.

Absolute KGE || Absolute ViT This model concatenates the absolute KGE (abs-KGE) of original dimension of 128, and the absolute ViT (Abs-ViT) of original dimension of 768. This leads to a hybrid vector of dimension 896. The first linear layer transforms entity embeddings into a 512-dimensional space.

Relative KGE || Relative ViT This model concatenates the relative KGE (rel-KGE) of dimension 700, and the relative ViT (Rel-ViT) of dimension 700. This leads to

a hybrid vector of dimension 1400. The first linear layer transforms entity embeddings into a 512-dimensional space.

Relative KGE || Absolute ViT This model concatenates the relative KGE (rel-KGE) of dimension 700, and the absolute ViT (Abs-ViT) of dimension 768. This leads to a hybrid vector of dimension 1468. The first linear layer transforms entity embeddings into a 512-dimensional space.

Relative KGE $\odot$ Relative ViT In this model, we combine the relative KGE (rel-KGE) and relative ViT (Rel-ViT) vectors using element-wise multiplication (Hadamard product as in $(A \circ B)_{ij} = (A \odot B)_{ij} = (A)_{ij}(B)_{ij}$). This operation helps us highlight the degree of similarity between each image and 700 selected anchors, considering both spatial (ViT) and semantic (KGE) information. The result is a hybrid 700-dimensional vector. The first linear layer transforms the embeddings into a 512-dimensional space.

IV.3.3.4 Evaluation Metrics

We employ identical performance metrics and maintain consistent training and testing data splits as used in Chapters II.2 and III.2.

IV.3.3.5 Interpretability Approach

Our aim was to understand how different embedding methods represent test images and discover relevant similarities with training instances.

Absolute Embeddings: Top Similar Image Analysis For both absolute KGE and absolute ViT embeddings, we utilize the embeddings of a test image and the training images to identify the top 5 most similar embeddings.

Relative Embeddings: Top Similar Anchors Analysis With relative representations for both KGE and ViT, we identify the top k similar anchors to a test image. This analysis involves three steps:

- **Top similar anchors based on Relative KGE embeddings (rel-KGE):** We find the top k anchors that are most similar to the test image based on the relative KGE embeddings.
- **Top similar anchors based on Relative ViT embeddings (Rel-ViT):** We identify the top k anchors with the highest similarity to the test image using the relative ViT embeddings.
- **Top similar anchors based on Hybrid embeddings (Hadamard Product):** We extract the top k similar anchors based on the composite representations created through the Hadamard product.

SPARQL Query the AKG To delve deeper into the knowledge learned by the model, we conduct a SPARQL query for each of the top images on the ARTstract-KG to uncover shared triples or common nodes. These shared triples may reveal that a significant portion of the top k most similar anchors share attributes such as certain objects, colors, etc. This process empowers us to unveil and interpret the insights the model has gained regarding the image's content and the shared characteristics among similar images.

```
PREFIX : <https://w3id.org/situannotate#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?s ?p ?o (COUNT(?img) AS ?sharedBy)
WHERE {
    VALUES ?img {
        :Image_ID_1
        :Image_ID_2
        :Image_ID_3
        :Image_ID_4
        :Image_ID_5 }
    {
        ?img ?p ?o .
    }
}
GROUP BY ?s ?p ?o
ORDER BY DESC(?sharedBy) ?s ?p ?o
```

IV.3.4 Results

IV.3.4.1 The ARTstact-KG

The ARTstact-KG represents a resource that systematically captures and formalizes intricate relationships within the ARTstract dataset. Comprising over 1.9 million triples, it encompasses data from more than 14,000 unique images and offers a profound understanding of perceptual semantics. The heart of the ARTstact-KG lies in the reification of annotation situations and perceptual semantic labels. Annotation situations capture details like geographical locations, timestamps, annotators, model architectures, and datasets. Similarly, perceptual semantic labels assigned to ARTstract images are reified as instances of the `Annotation` class, forming a complex network of connections, linking each annotation to its corresponding annotation situation, the associated image, the lexical entry used, the assigned annotation strength, the annotation role, and the ConceptNet concept that typifies it. FigureIV.3.9 provides an example of the wealth of

information within the ARTstact-KG, as visualized through a single annotation seen via the Protégé tool.[4] As shown in Figure IV.3.10, ARTstact-KG is further enriched with high-level linguistic frames extracted from image captions. These frames are extracted using FRED and WordNet synsets as triggers, offering a comprehensive linguistic context for each image. This enrichment enhances the knowledge graph's expressiveness and provides a deeper understanding of the linguistic context associated with each image.

IV.3.4.2 AC Image Classification Performances

Input	Embedding Type	D	Macro F1
KGE-only	Absolute	128	0.22
	Relative	128	**0.27**
ViT-Only	Absolute	768	**0.30**
	Relative	700	0.28
KGE || ViT	Absolute KGE || Absolute ViT	896	0.31
	Relative KGE || Relative ViT	1400	**0.33**
	Relative KGE || Absolute ViT	1468	**0.32**
Relative KGE $\odot$ Relative ViT		700	0.29

Table IV.3.2: Macro F1 scores using each of the embeddings as inputs, trained for 50 epochs. D: Dimensionality; ||: concatenation; $\odot$: Hadamard product.

In our study, we explored the utilization of ARTstract KG embeddings for the task of AC image classification. The results of our experiments provide valuable insights into the effectiveness of different approaches. We present the results of our experiments in Table IV.3.2, also visualized in Figure IV.3.5.

Absolute versus Relative Embeddings Our results show that the absolute (original) KGE embeddings (Abs-KGE) achieved a Macro F1 score of 0.22, while their relative KGE counterparts (Rel-KGE) outperformed the absolute ones with a Macro F1 score of 0.27. This suggests that the relative representation approach significantly enhances Knowledge Graph Embeddings' performance in AC image classification. For deep feature embeddings (ViT), the absolute version (Abs-ViT) scored 0.3 in Macro F1, while the relative counterparts (Rel-ViT) scored slightly lower at 0.28. These results indicate that applying the relative representation approach may lead to subtle performance degradation in ViT embeddings. Our findings highlight performance disparities between absolute and relative embeddings, depending on the original embedding method.

Hybrid Embeddings We explored the effectiveness of hybrid approaches combining ARTstract KG embeddings with ViT deep feature vectors for AC image classification.

[4] https://protege.stanford.edu/. Access date: December 2023.

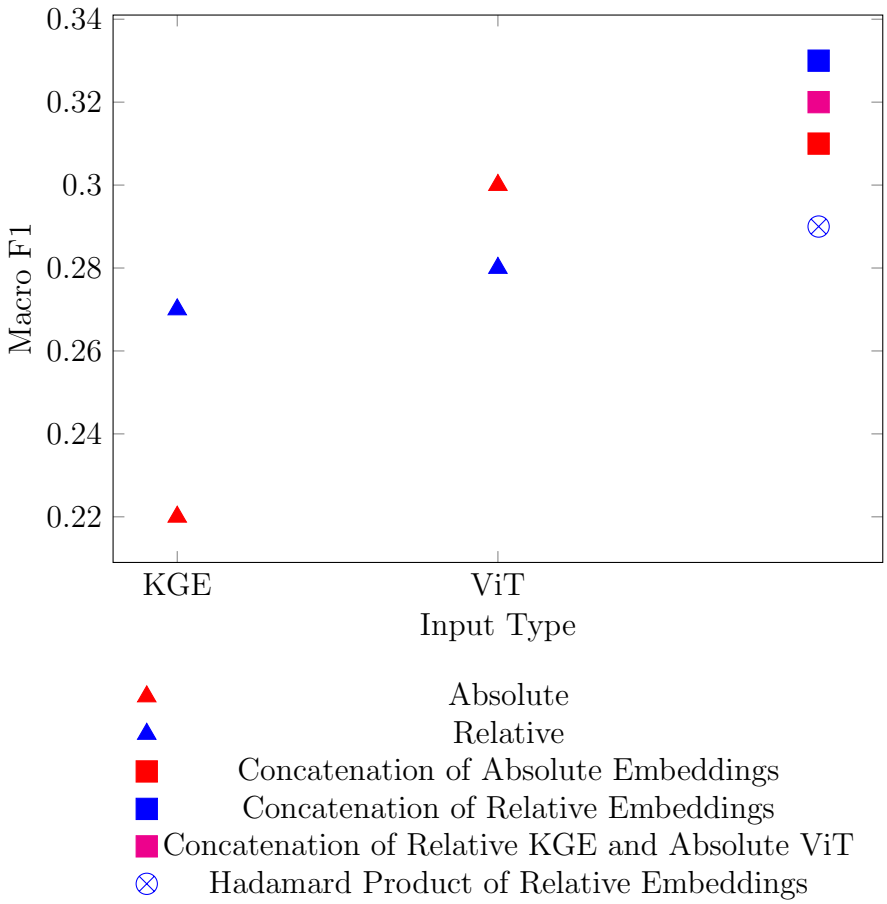


Figure IV.3.5: Performance (Macro F1) for different input embeddings

Table IV.3.2 showcases the accuracy achieved by different hybrid methods, with the highest F1 score (**0.33**) attained through the concatenation of the relative embeddings (both relative KGE and relative ViT) underscoring their synergistic effect on classification accuracy. The combination of relative KGE and absolute ViT embeddings achieved the second-highest F1 score (0.32), while the concatenation of the absolute embeddings (both absolute KGE and absolute ViT) achieved the third-best score, 0.31. These three methods outperformed the use of only one of the embeddings, the best of which had been absolute ViT with an F1 score of 0.30. Additionally, the Hadamard product of relative embeddings obtained an F1 score of 0.29, surpassing KGE-only methods, and achieving comparable performance to absolute ViT.

Comparison with State of the Art

In Table IV.3.3, we provide a comparative analysis of our models, which combine ARTstract KG embeddings (KGE) with ViT deep feature vectors for AC image classification.

Method	Scores	Paradigm
	Macro F1	
XGB	0.20	PS
SVM	0.20	PS
Absolute KGE	0.22	SPK
VGG-16	0.23	DL
Naive Bayes	0.24	PS
ResNet-50	0.24	DL
Relative KGE	0.27	SPK
Relative ViT	0.28	DL
Relative KGE $\odot$ Relative ViT	0.29	Hybrid
Absolute ViT	0.30	DL
Absolute KGE || Absolute ViT	**0.31**	Hybrid
Relative KGE || Absolute ViT	**0.32**	Hybrid
Relative KGE || Relative ViT	***0.33***	Hybrid

Table IV.3.3: Comparative analysis of the KGE-based models along with the best performing classical ML and DL models. The top-performing model is highlighted in both bold and italics. The second-best performing models are denoted in bold.

We benchmark these models against *end-to-end vision* deep models (Chapter II.2) and the top-performing classical machine learning (ML) models from our *perceptual semantics* paradigm (Chapter III.2). Among the classical ML models, XGBoost (XGB) and Support Vector Machines (SVM) had achieved macro F1 scores of 0.20, while the Naive Bayes model reached a score of 0.24. For the DL models, ResNet-50 and VGG-16 attained macro F1 scores of 0.24 and 0.23, respectively. In this contest, our KGE-only based models exhibited significant potential, considering their lack of access to pixel-level features. The Absolute KGE model achieved an F1 score of 0.22, while the Relative KGE model demonstrated an even more remarkable score of 0.27, outperforming all CNN and ML methods.

The hybrid approaches we tested delivered the highest performance (Figure IV.3.11), surpassing all other methods encountered in this research and in the state of the art. With the Relative KGE || Relative ViT approach, the concatenation of the two relative embeddings yielded an outstanding F1 score of 0.33, which is the highest score in this task to our knowledge. The combination of Relative KGE and Absolute ViT embeddings achieved the second-highest performance in this work, with an F1 score of 0.32. The concatenation of Absolute KGE and Absolute ViT embeddings attained an F1 score of 0.31. Our last hybrid method, the Hadamard product of the two relative embeddings, performed slightly worse than the Absolute ViT (0.29 versus 0.30) but is comparable and more interpretable. These outcomes suggest that combining KGE-based perceptual

knowledge with ViT embeddings holds promise for enhancing the accuracy of AC image classification, highlighting the complementarity of these embedding types.

IV.3.4.3 Interpretability Results

Absolute ViT vs Absolute KGE: Training Images Similarity

In this section, we delve into the comparative analysis of two resulting embeddings from processing images via the Absolute ViT and Absolute KGE paradigms.

Perceptual Disparities Our results demonstrate notable disparities when utilizing the two absolute embeddings to retrieve the top 5 most similar training images for a test image. To illustrate the divergent behavior of these models, we present the outcomes of three test images in Figure IV.3.6. For the top test image, Image 14817, Absolute ViT's top similar images prominently feature the United States flag, with 3 out of 5 being flag images and 4 out of 5 being tagged with the ground truth *freedom*. This observation indicates that ViT places significant emphasis on the United States flag's presence, possibly revealing a bias in its training data. The appearance of stars against the blue background is notably salient in ViT's assessment. In contrast, the most similar images generated by Absolute KGE embeddings are labeled with the *comfort* ground truth and demonstrate a stronger visual and semantic connection with the lower portion of the test image, encompassing elements like grass, fields, trees, and greens. These findings suggest that the two embeddings, in specific instances, capture distinct types of information. In certain cases, not only do the two embeddings capture different aspects of an image, but one also outperforms the other in terms of semantics. For instance, in the case of the bottom test image shown in Figure IV.3.6, Absolute KGE demonstrates superior semantic performance. It successfully associates the Statue of Liberty with anchors that not only align with the correct ground truth but also prominently feature the Statue of Liberty. In contrast, Absolute ViT fails to find matching anchors related to the ground truth, resulting in less semantically coherent results. Conversely, for the third test images in Figure IV.3.6, we observe a scenario where Absolute ViT surpasses KGE in terms of semantics. For example, in this case, ViT effectively highlights the significance of the combination of a horse and wheels, implying the presence of a powered vehicle, while KGE appears to primarily focus on the horse.

High-Level Semantic Proficiency In addition to the observed disparities between Absolute ViT and Absolute KGE, our investigation reveals that even when both individual absolute embeddings make correct predictions, they demonstrate distinct understandings of images. Specifically, it appears that KGE excels in identifying higher-level semantics. To illustrate this phenomenon, we present results for two images tagged with the ground truth *comfort*, as depicted in Figure IV.3.7. For the test image on top, Absolute ViT excels in encoding what could be termed the images' "aesthetics," emphasizing elements such as colors and artistic composition. Absolute KGE, instead, stands out

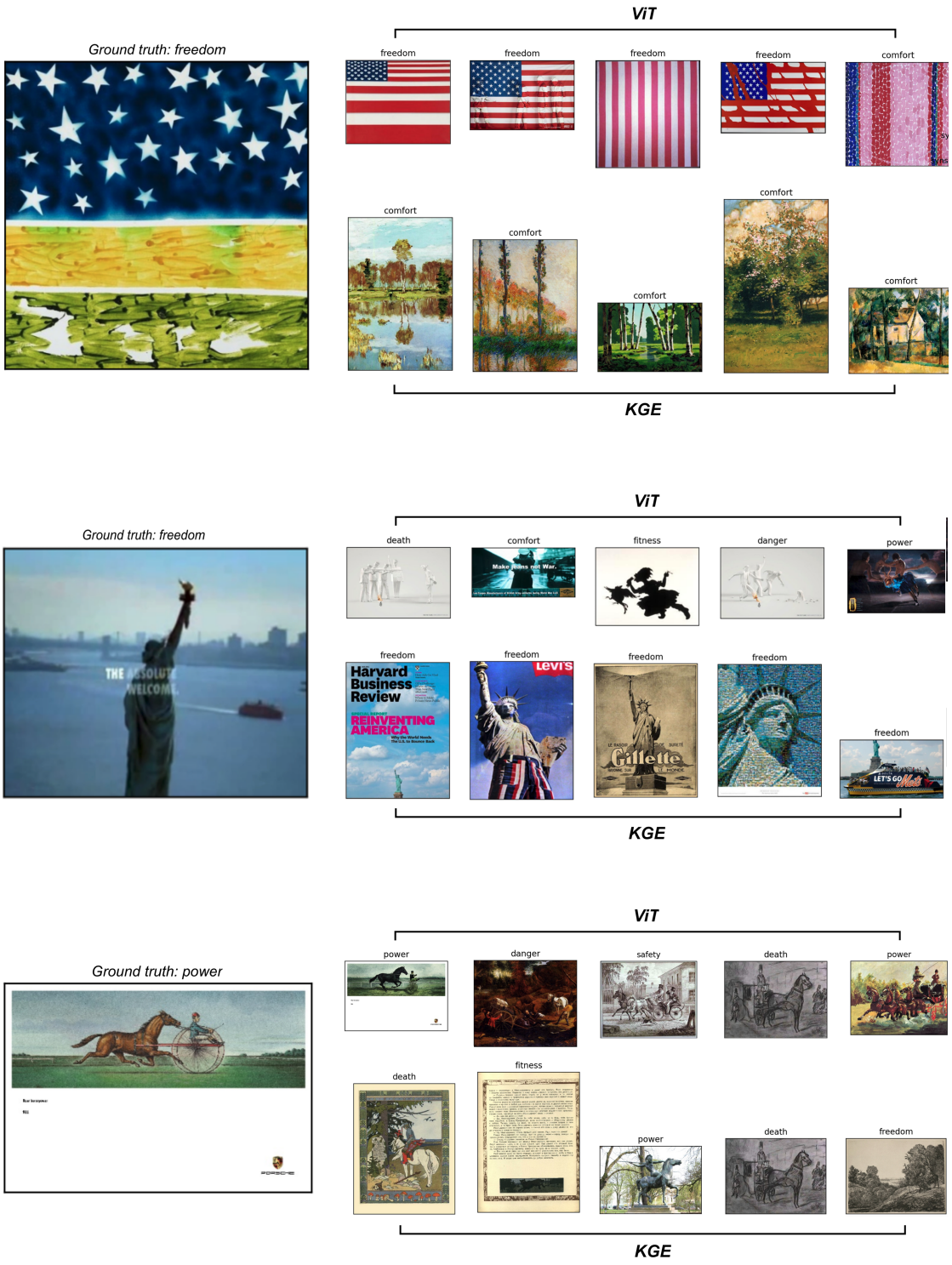


Figure IV.3.6: Absolute ViT vs. Absolute KGE embeddings capture different aspects of ARTstract images. Top: Absolute ViT captures aspects that resemble the United States flag while KGE captures more landscape-related features, Middle: Absolute KGE demonstrates superior semantic performance than ViT by encoding similarities with perceptually diverse visions of the Statue of Liberty; Bottom: Absolute ViT encodes similarities between images that make more semantic sense for the *power* ground truth.

by recognizing the explicit semantics within the image, representing a woman comfortably reading. KGE achieves this by matching all the top 5 similar images to images of women reading (See Figure IV.3.7, bottom row), while ViT fails to match the test image with any training images exhibiting the same semantics of the depicted scene. Similar observations apply to the test image at the bottom of Figure IV.3.7. Both Absolute ViT and Absolute KGE accurately labeled the test image. Still, KGE demonstrates a superior ability to understand the connection between comfort and the act of sitting on a couch. This example underscores the distinction between the two models: while both get the prediction right, Absolute KGE's most similar images are more semantically related, featuring women depicted doing specific actions, which is not the case with ViT's most similar images. Even though both models correctly predict the ground truth, KGE leverages perceptual semantics, likely detecting the presence of a couch as a key element in establishing similarity. In the Discussion Section (Section IV.3.5), we present other intriguing examples highlighting KGE's apparent superiority in capturing higher-level semantics compared to ViT, which tends to focus more on compositional and low-level features.

Relative Embeddings: Training Anchors Similarity

We implemented our interpretability approach for relative representation embeddings using prototype anchors. To discern the top 5 anchors to which each test instance bears the greatest similarity, we considered three different embeddings: relative KGE, relative ViT, and the hybrid relative (derived from the Hadamard product of relative ViT and relative KGE). Some exemplary results are presented in Figure IV.3.8. Within each subfigure, rows of images showcase the top 5 images with the most "embodied" similarity, as determined by relative ViT embeddings (top row), the top 5 images with the most "symbolic" similarity, as determined by relative KGE embeddings (middle rows), and the top 5 images with the most "embodied-symbolic" similarity, as determined by the Hadamard product of the two relative embeddings. Additionally, accompanying each row are frequently shared ARTstract-KG nodes for the set of images. These nodes were extracted through a SPARQL query on the knowledge graph, revealing shared attributes and characteristics contributing to the perceived similarity between images. Our goal in all instances is to extract meaningful insights from the models by analyzing common triples in the knowledge graph. This process aids us in comprehending the underlying factors that contribute to image similarity as learned by the models.

In Figure IV.3.8, we present results demonstrating the effectiveness of the hybrid approach, which combines both relative representations. For instance, in the top image of the figure with the *fitness* ground truth, both relative ViT and relative KGE embeddings independently exhibit high similarity to fitness-tagged anchors, with some overlap. Notably, the hybrid vector for the test image yields the top 5 anchors, all tagged with *fitness*, showing that the hybrid embedding excels in recognizing spatially-semantically similar anchors. In a similar context, the bottom example with the *danger* ground truth follows a comparable trend. Both relative ViT and relative KGE embeddings

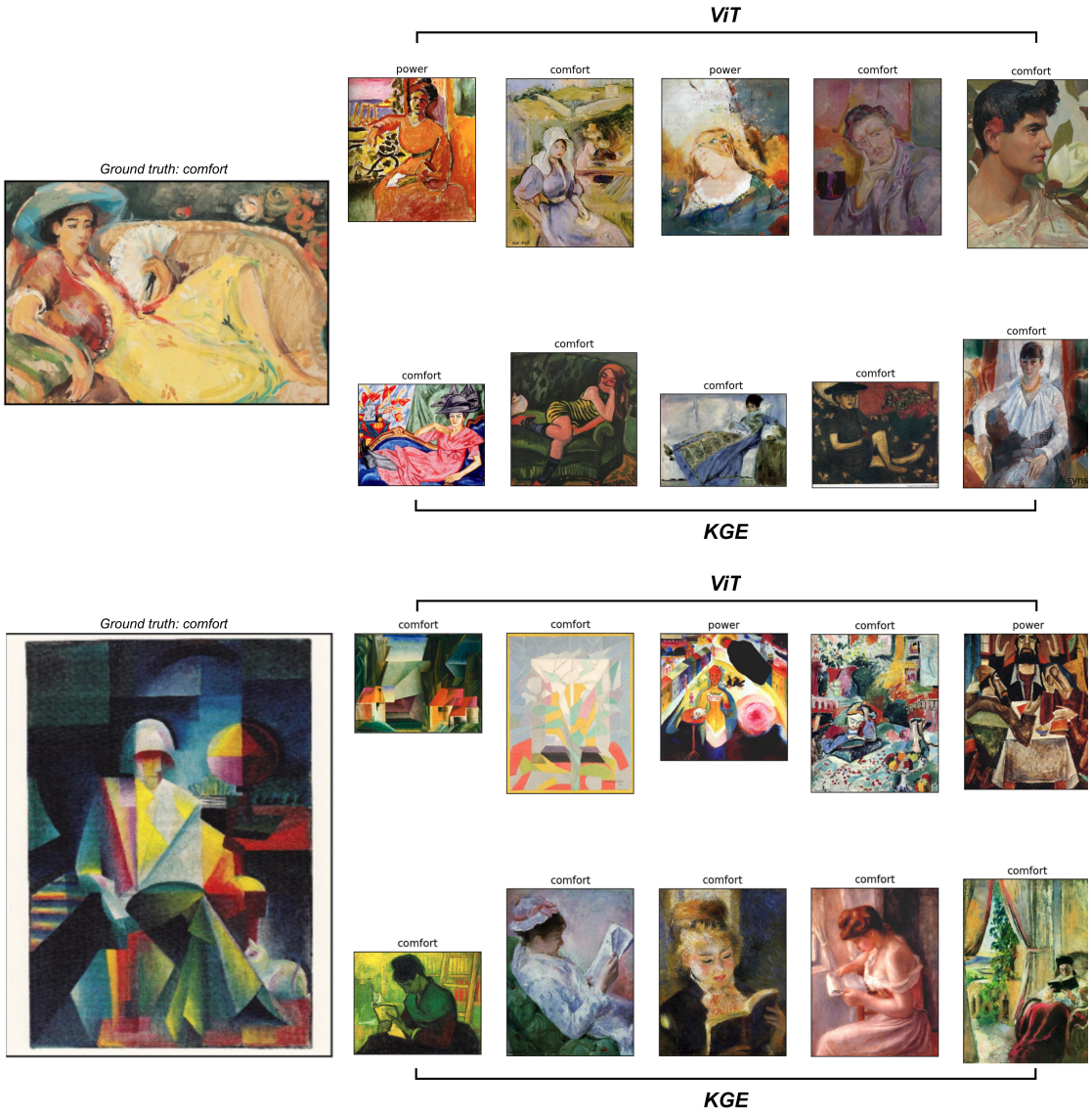


Figure IV.3.7: Contrasting semantic proficiency of Absolute KGE vs. Absolute ViT. The top image illustrates ViT's focus on colors and textures (aesthetics), whereas KGE excels in recognizing explicit semantics, particularly women sitting on couches. In the bottom image, KGE effectively encodes the semantics of reading a book in the test artwork.

independently identify the importance of water and boats among similar anchors but may also include some incorrect ground truth anchors. In contrast, the hybrid version consistently identifies all of the top 5 anchors with the correct ground truth, suggesting complementarity between these two relative representations.

IV.3.5 Discussion

IV.3.5.1 The ARTstract-KG

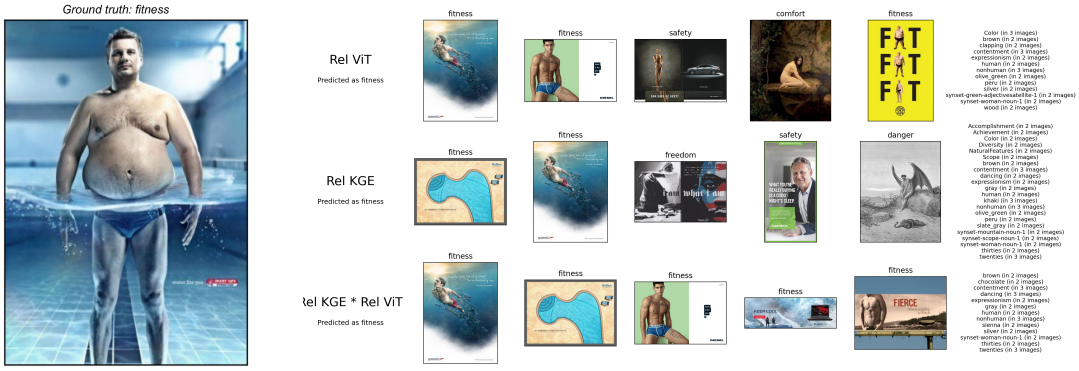
The ARTstract KG is a robust and context-aware structured knowledge repository of metadata annotations encompassing more than 14,000 cultural images. These annotations cover a spectrum of perceptual semantic aspects, painting a comprehensive picture of the images. They include details about the source dataset, the context in which they evoke concepts, automatically detected art styles, depicted objects, actions, dominant colors, the presence of human subjects, age tiers, emotions, and image captions. Moreover, the AKG provides insight into the specifics of when, where, and how these perceptual semantic annotations were made, as well as the annotation strength associated with each annotation. This contextual richness effectively makes these annotations *situated* ground truths. One of the noteworthy aspects of the AKG is the reification of perceptual semantics. This reification allows for connections to be established with commonsense knowledge sources like ConceptNet, enabling interpretable reasoning over perceptual semantic data. This facilitates the extraction of high-level linguistic frames, contributing to the linguistic understanding of the images. In this way, the AKG contains the formalization of both human- and machine-readable representations of images, advancing our capacity to query and question machine vision's automatic comprehension of visual content. Furthermore, the AKG serves as a repository for the annotations of ACs, further enhancing its value in the context of AC image classification. This knowledge graph serves as a valuable resource for comprehending contextual information in visual sensemaking, establishing a solid foundation for further research and analysis.

IV.3.5.2 AC Image Classification Performances

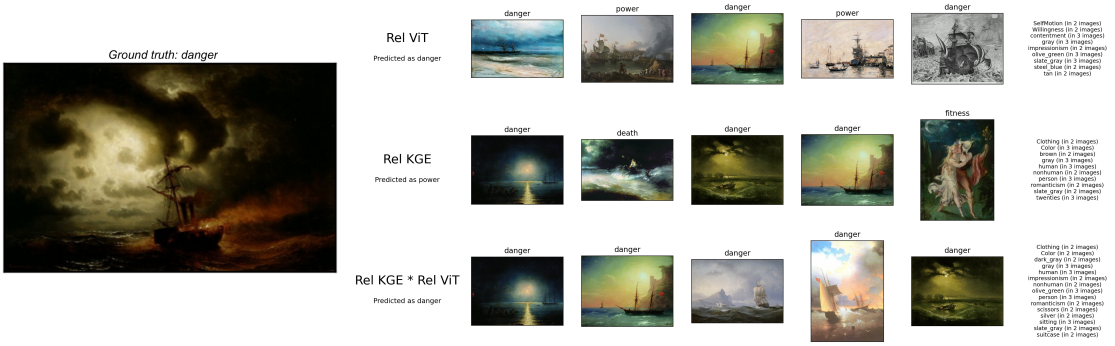
Absolute versus Relative Embeddings

In our study, we tested the relative representation method introduced by Moschella et al. (2022) [261], wherein each instance is encoded in relation to selected anchor points. For this purpose, we selected 100 anchors for each of the 7 target classes, ultimately representing images in relation to these 700 anchors. Consequently, this representation may be perceived as a more 'semantic' representation, as each embedding becomes semantically biased towards a balanced representation of how each instance relates to the seven target clusters.

The results of our study indicate that while the initial performance of the KGE approach was lower than that of Vision Transformer (ViT), the incorporation of the



(a) Interpretability results for a test image with ground truth *fitness*. The hybrid embedding leverages complementary information from both relative embeddings to encode the highest similarity of the test image with all training anchors tagged as *fitness*.



(b) Interpretability results for a test image with ground truth *danger*. The hybrid embedding exhibits similarities with anchors exclusively tagged as *danger*, with some similarities predominantly originating from the relative ViT and others from the relative KGE embeddings.

Figure IV.3.8: Two representative examples of our interpretability approach with relative representations. We identify the top anchors to which each test instance bears the most similarity. In each subfigure, rows of images show the top 5 "sensory-perceptually" (most similar Rel-ViT embeddings), "symbolically" (most similar rel-KGE embeddings), and hybridally similar images (most similar Rel-ViT $\odot$ rel-KGE hybrid embeddings (Hadamard product embedding). Alongside each row, we display frequently shared ARTstract-KG nodes, obtained through a SPARQL query on the KG.

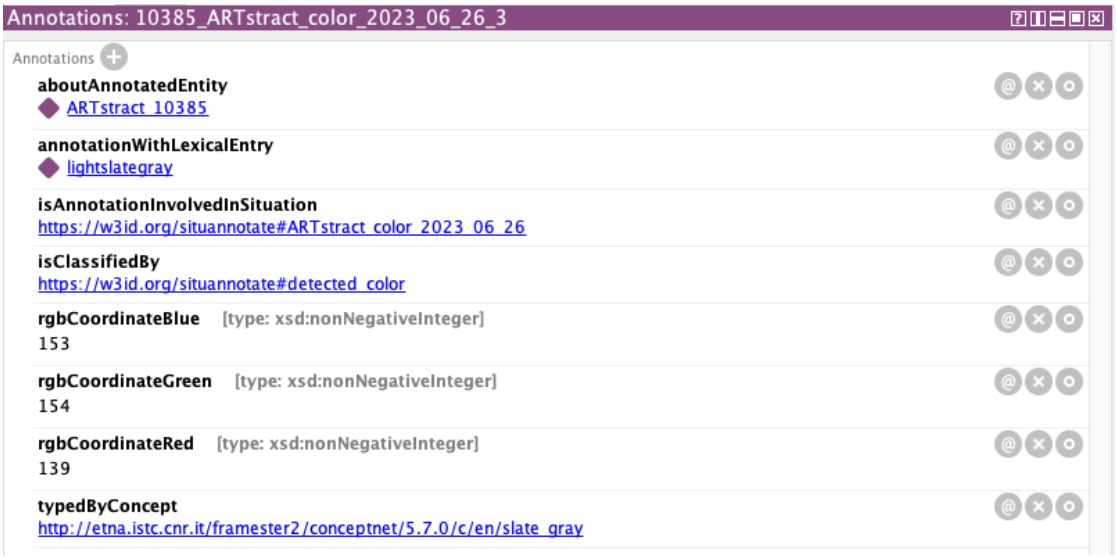


Figure IV.3.9: Protégé snapshot of ARTstract KG triples about an annotation.

relative representation approach significantly improved KGE's classification capabilities. Absolute KGE initially achieved a Macro F1 score of 0.22, which was outperformed by Absolute ViT at 0.30. However, with the introduction of the relative representation method, Relative KGE closed the performance gap, achieving a Macro F1 score of 0.27, which was now competitive with ViT. These findings suggest that the relative representation method significantly enhances the performance of KGE-based models by providing more meaningful cluster-level semantic information. The superior performance of Rel-KGE can be interpreted as an indication that this method introduces a more 'semantic' representation, consequently elevating the semantic resolution. Therefore, these findings underscore the potential of the relative representation method to empower KGE-based image classification, making it a valuable alternative to ViT. It suggests that the enhanced semantic representation introduced by the relative approach provides KGE with a substantial boost in its classification performance.

Conversely, in the context of ViT embeddings, the results revealed a different trend. Absolute ViT (Abs-ViT) embeddings achieved a higher Macro F1 score compared to Relative ViT (Rel-ViT) embeddings. This implies that ViT, designed to handle pixel-level information, may not benefit from the semantic bias introduced by the relative representation approach. The lower performance of Rel-ViT implies a potential loss of fine-grained local differences and similarities between images, which are critical for capturing spatial resolution. This suggests that relative representations may not be suitable for pixel-level models like ViT, potentially leading to performance issues.

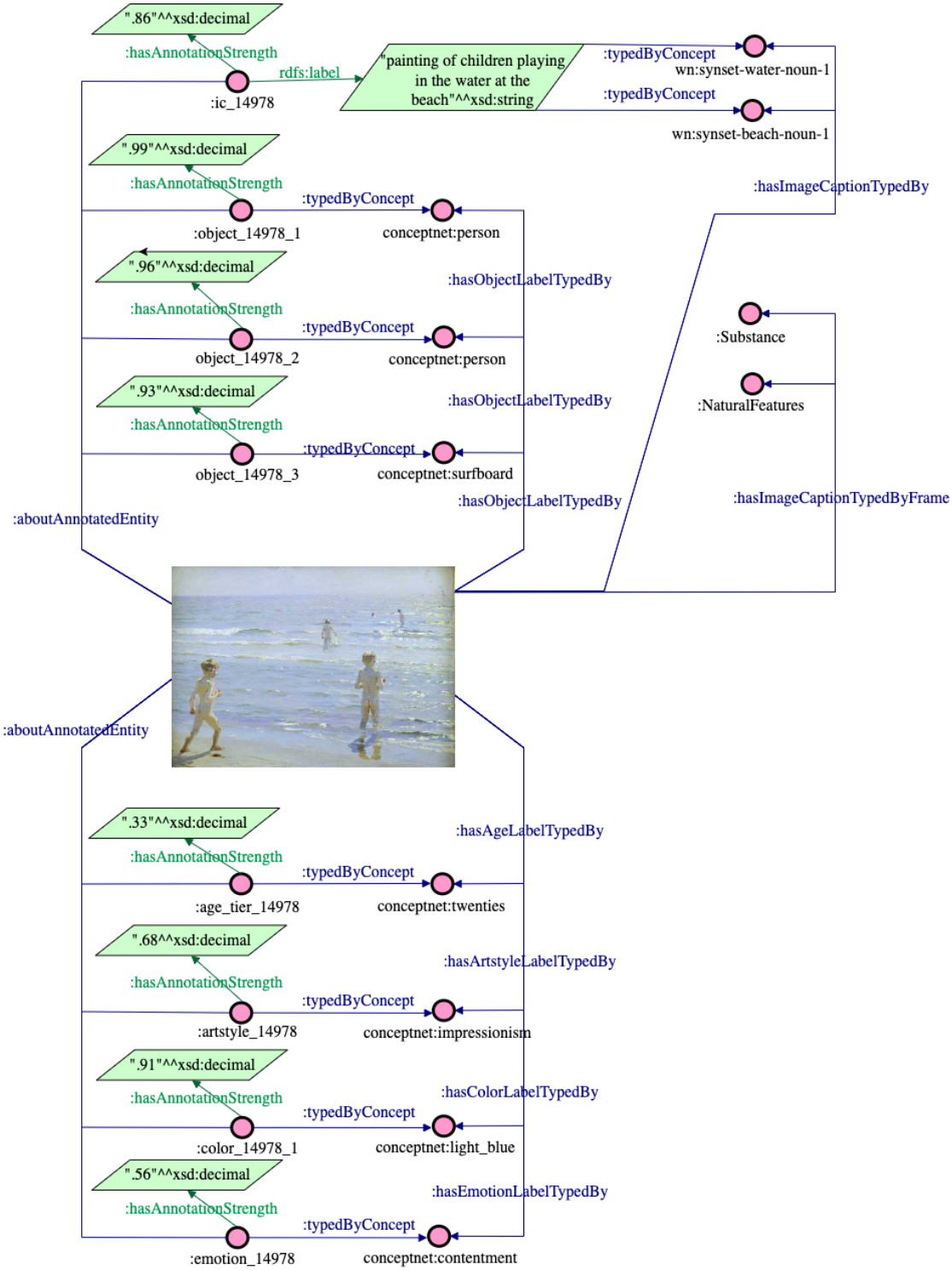


Figure IV.3.10: Subset of the A-Box of ARTstract-KG, showing the types of commonsense linguistic knowledge connected to a single image instance. Most annotations are typed by ConceptNet concepts, while the image captions are typed by WordNet concepts as well as by linguistic frames.

Hybrid Approaches

In our investigation, we experimented with combining KGE and ViT embeddings to assess whether their joint usage could enhance performance compared to using either type in isolation as input for the MLP classifier. As seen in Figure IV.3.5, the top F1 score of 0.33 was achieved through the *concatenation of relative embeddings*. This finding signifies the effectiveness of combining relative KGE and relative ViT embeddings in bolstering classification accuracy. Relative embeddings provide valuable perspectives on image relationships and shared attributes, thus contributing substantially to the model's overall accuracy. In comparison, the F1 score for combined absolute embeddings aligns closely with the performance of the absolute ViT by itself, scoring 0.31 and 0.30, respectively. This implies that pairing two relative embeddings yields more substantial performance improvements than uniting two absolute embeddings, highlighting the complementary nature of relative embeddings. The second-highest F1 score, at 0.32, was achieved by the hybrid approach concatenating the best-performing individual embeddings (relative KGE and absolute ViT). This result highlights the complementarity of the relative KGE and absolute ViT embeddings, emphasizing the significant impact of their combination on the model's overall performance.

Our results also draw attention to the Hybrid Approach utilizing the Hadamard product, which achieved an F1 score of 0.29. While surpassing the KGE-only-based approaches, it slightly lags behind the performance of absolute ViT. Nonetheless, it outperforms both relative KGE and relative ViT embeddings. We specifically opted for the Hadamard product ($A \circ B$) because it emphasizes the degree of similarity between each image embedding and the 700 selected anchors, taking into account spatial (ViT) and semantic (KGE) information. This operation accentuates anchors that exhibit the highest similarity to a given image when assessed from both spatial and semantic perspectives. Conversely, features with low scores in one of the vectors will lead to lower values in the resulting hybrid vector. This approach proves valuable in identifying anchor images that share pronounced similarities with a given anchor but only in one of the two embedding spaces. This ability to unveil unique characteristics captured by each modality underscores the Hadamard product's effectiveness in pinpointing common features and characteristics, facilitating the identification of anchors that hold particular significance and dual-mode similarity to the image of interest.

Comparison with State of the Art

The comparison with classical ML and DL models reaffirms the efficacy of our hybrid approaches, which outperform all the other methods (see Figure IV.3.11), providing evidence that the integration of ARTstract KG embeddings with ViT deep feature vectors unlocks substantial improvements in image classification accuracy. While classical models handling the unsituated perceptual semantics have demonstrated performance comparable to CNN methods in Chapter III.2, it is the synergy between the situated perceptual knowledge encoded in the relative KGE embeddings with the sensory-perceptual understanding captured by the ViT embeddings that leads to the highest F1 score of 0.33.

These findings collectively underscore the significant impact of our hybrid approaches, with the Relative KGE || Relative ViT model emerging as the top-performing and most promising method, effectively enhancing the precision of AC image classification.

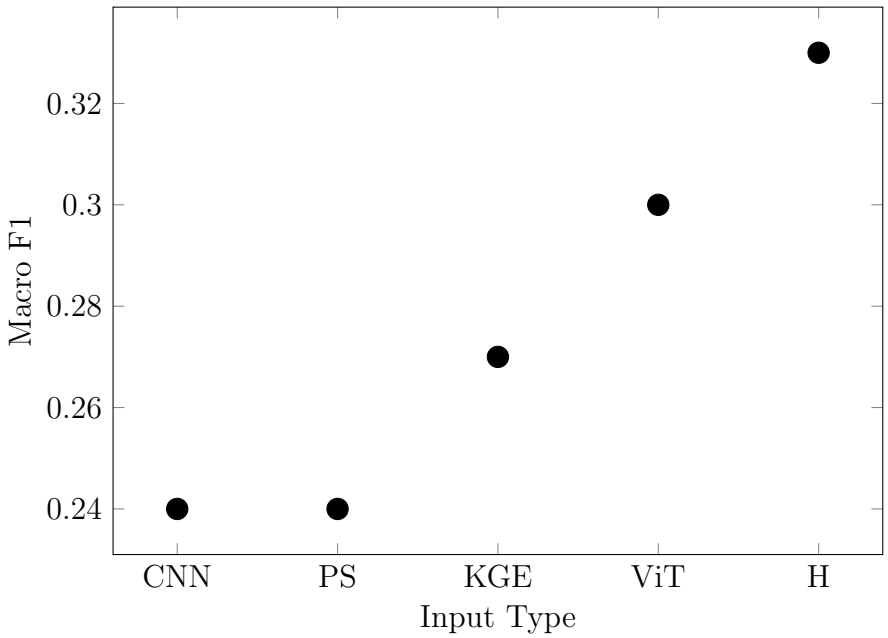


Figure IV.3.11: Our hybrid approaches have the highest known performance for the task of AC image classification. Best Macro F1 Scores for different input embeddings. PS: Naive-Bayes; CNN: ResNet-50; KGE: Relative TransE KGE; ViT: Absolute ViT; H: Concatenation of Relative KGE and Relative ViT.

IV.3.5.3 Interpretability

These interpretability experiments aid us in comprehending the inner workings of our models and can guide further improvements in image classification and retrieval tasks. They highlight the models' ability to capture both symbolic and embodied aspects of image content, contributing to a more holistic understanding of image similarity.

Similar Images: Absolute Embeddings Comparison

Perceptual Disparities: ViT and KGE Comparative Analysis Our results underscore the presence of significant disparities between the behaviors of Absolute ViT and Absolute KGE when processing and interpreting images. In particular, these disparities may hint at biases inherent in the training data and architectural differences between the models. For example, the observation that Absolute ViT places substantial emphasis on the United States flag in one test image (see Figure IV.3.6) could be

indicative of a bias in its training data. This bias may have resulted from an overrepresentation of United Statesian content and flag-related images or certain visual patterns in the training dataset, leading ViT to exhibit this behavior. It is essential to be aware of such biases, as they can impact the model's generalizability and performance on diverse tasks. On the other hand, Absolute KGE's performance, especially being biased towards similar images with *comfort* as ground truth, may be attributed to its reliance on only ARTstract for training, which is strongly unbalanced and thus biased towards the comfort cluster. Furthermore, we observe instances where Absolute KGE and Absolute ViT excel differently in semantics, illustrating the nuanced capabilities of each model. These differences might stem from their unique architectures and training objectives. Understanding these variations in interpretability and performance is crucial when choosing between the two embeddings for specific applications.

Semantic Proficiency: KGE's Edge Our results raise intriguing questions about the interpretability and semantic understanding of images by Absolute ViT and Absolute KGE. The disparities observed in how these models handle images, even when both achieve accurate predictions, suggest the presence of underlying biases and architectural differences. In the case of the top image of Figure IV.3.7, where Absolute ViT emphasizes aesthetics and KGE identifies explicit semantics, it becomes apparent that ViT may exhibit a bias towards certain visual patterns and color schemes. These biases may be reflective of the data distribution in its training dataset and the way it has learned to interpret comfort-related images. Conversely, KGE's ability to focus on explicit semantic elements suggests that it may be more adept at capturing high-level concepts in images. The bottom test image highlights how both models can make accurate predictions, yet KGE excels in understanding the nuanced relationship between comfort and sitting on a couch. This suggests that KGE may incorporate higher-level semantics in its representations, potentially through object detection or pattern recognition.

We identified other cases showcasing the KGE method's superiority at capturing higher-level semantics than ViT. In Figure IV.3.12, the test image portrays two individuals in an intimate setting. ViT-based similar images primarily focus on pixel-level resemblances related to dark colors and textures. In contrast, KGE demonstrates a superior understanding of the scene, emphasizing the presence of multiple individuals engaged in intimate interactions. While most ViT-similar images depict single individuals, the majority of KGE-generated similar images depict scenes involving two or more people in intimate settings. This showcases a scenario where KGE excels in capturing high-level semantics, which ViT may overlook because of its emphasis on compositional and low-level features. While ViT may excel in recognizing detailed visual features, KGE appears to have an edge when it comes to interpreting scenes, especially those involving complex interactions and higher-level semantics. The ability of KGE to understand the presence of multiple individuals in an intimate setting suggests its potential for tasks that require interpreting social interactions, relationships, or other complex high-level visual cues.

In Figure IV.3.13, we present another compelling example that highlights the KG

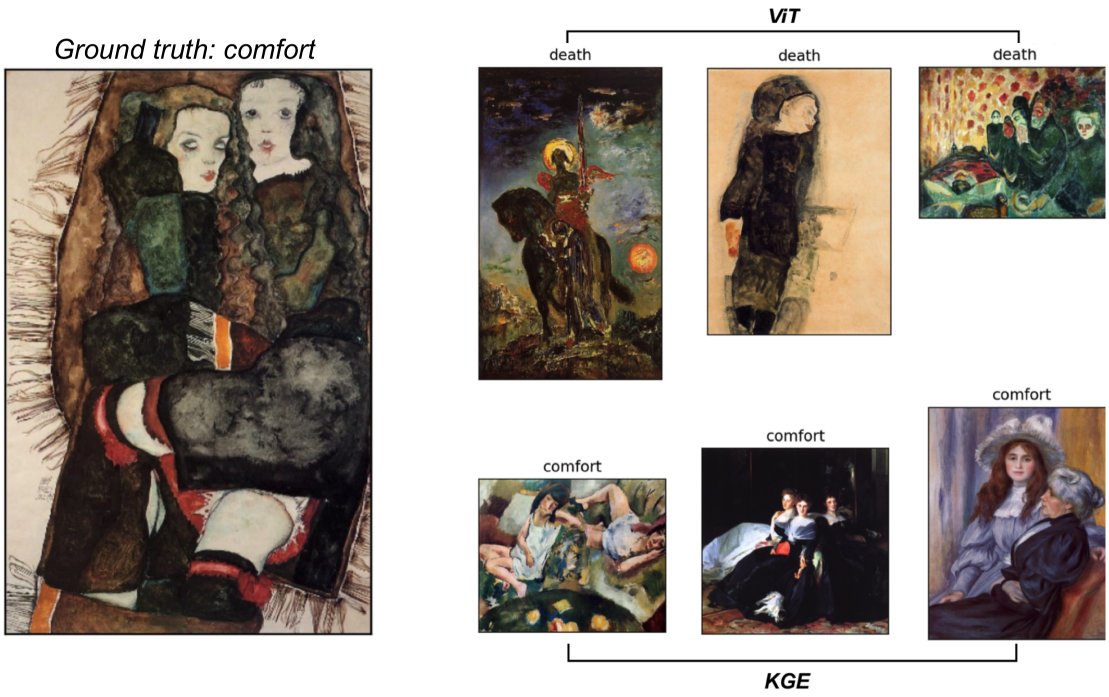


Figure IV.3.12: ViT focuses on colors and textures, while KGE excels in understanding complex scenes, emphasizing multiple individuals in intimate interactions.

methods' superior ability to capture perceptual semantics. The test image portrays a bed with a comforter. This case illustrates the challenges faced by ViT, which seems to get confused with sensory details like colors and lines, rendering it unable to abstract the underlying semantic content. In contrast, KGE exhibits a strong capacity to connect the visual content to the broader semantics of a bed, as evidenced by the most similar images it generates. This case exemplifies the contrast between ViT and KGE in terms of their perceptual understanding and semantic abstraction. ViT's focus on sensory features, such as colors and lines, limits its ability to identify the broader semantic context, often leading to confusion in the task. In this specific case, ViT may struggle to differentiate the test image from other visually similar patterns or textures. Conversely, KGE's proficiency in connecting the image to the concept of a bed indicates its capability to grasp high-level semantics, even when presented with visually complex images.

Figure IV.3.14 presents a particularly intriguing case where the KG-based methods demonstrate their capacity to capture high-level semantic concepts effectively. The test image in question is categorized under the ground truth label "death" likely due to its depiction of a convoy of vehicles resembling ambulances, akin to those dispatched to war zones. In this instance, the images retrieved as most similar by ViT are predominantly tagged with *comfort*. This discrepancy likely arises from ViT's focus on color composition, warm tones, and textural elements present in both the test image and the retrieved

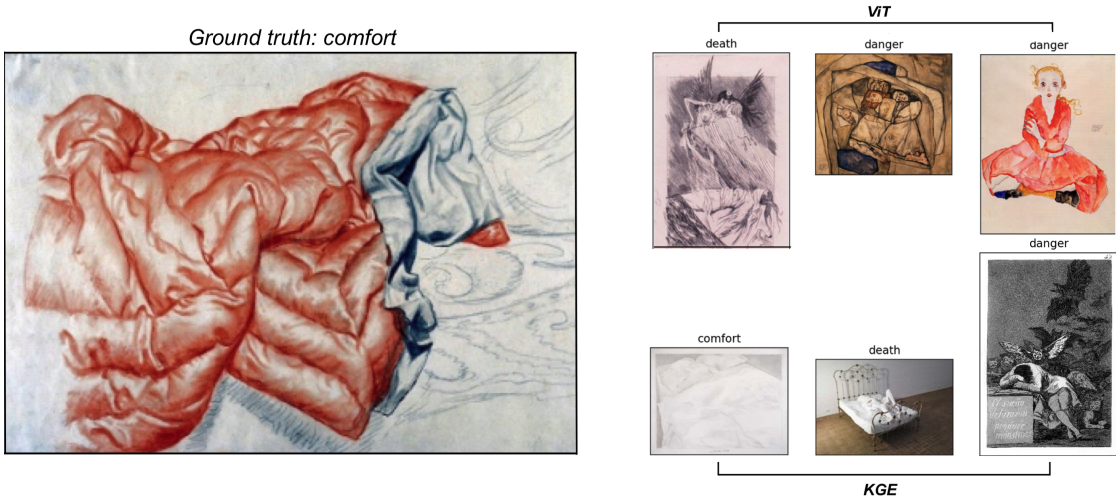


Figure IV.3.13: ViT struggles with sensory details, while KGE effectively connects the image to the broader concept of a bed.

images. The majority of the top images retrieved by ViT are set in outdoor scenes featuring open spaces, buildings, and natural landscapes. While these elements may be visually related to the test image, they are largely irrelevant to the ground truth. In contrast, the top three most similar images based on the KGE embeddings, although not closely related in terms of concrete visual elements, all share the correct ground truth of *death*, primarily evoking it through the presence of crosses and crucifixion imagery. This indicates that the KGE model has successfully linked images featuring crosses with the concept of death, a connection that takes precedence over the presence of open spaces and natural settings associated with *comfort* images (the ViT misclassifications).

Overall, when compared with relative-ViT, relative KGE representations select anchors more closely aligned with the semantic content of the target image. This proficiency of relative KGE embeddings is particularly noteworthy, considering the context in which the ARTstract-KG was constructed. This context involved automated (non-human evaluated) perceptual semantics detection, without manual semantic coherence checks, introducing inherent noise compounded by the complexities of cultural art images, which often lack discrete objects and other detected categories. Despite these challenges, our qualitative analysis of relative KG embeddings highlights the capacity of KGE embeddings to implicitly encode essential high-level semantics, a pivotal element in our study.

As such, despite our initial observation of ViT's superior F1 performance, which suggested a more "grounded" representation of ACs within its latent space, our interpretability experiments reveal a discrepancy. ViT's ability to capture semantic content at a level as high as the KGE representation, which makes sense to humans, falls short. We believe that this discrepancy is primarily a consequence of the prototype selection

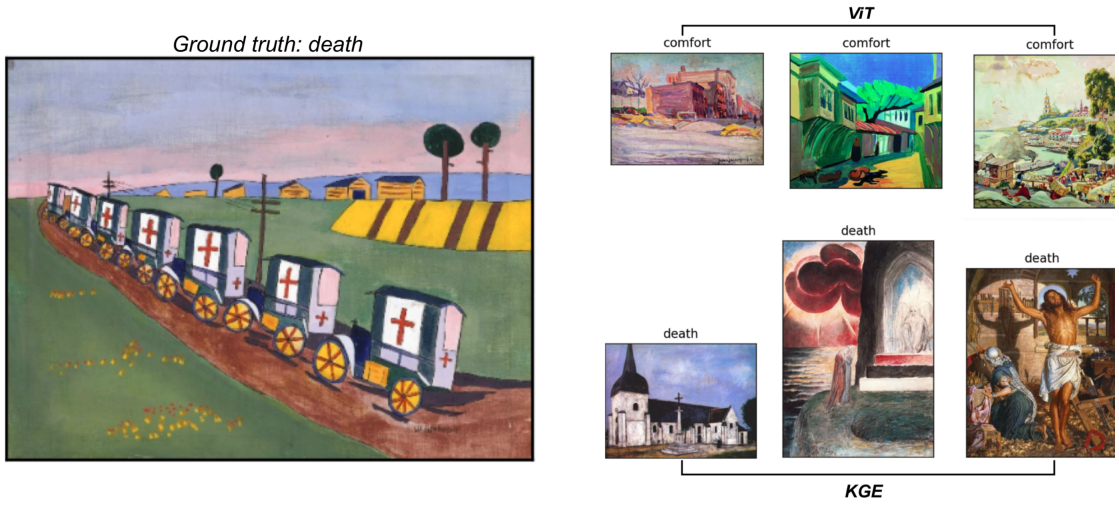


Figure IV.3.14: ViT misclassifies as *comfort*, but KGE successfully associates images with crosses to the concept of *death*.

process, where images are represented based on their similarity to these prototypes. Essentially, ViT's latent space heavily relies on the noise accumulated from its extensive training dataset. However, when we transform this deep representation into a relative form, we introduce a strong prior assumption, expecting images that evoke the same AC to exhibit semantic similarity. This transformation does not perfectly align with ViT's latent space; instead, it confines the representation to specific regions within that space. This constraint potentially limits ViT's ability to express semantic relationships, as it can no longer rely solely on pixel-wise perceptual features but must effectively position images within its latent space. As a result, the images obtained in this process may seem perplexing because the model's internal representation significantly differs from human perception. It primarily depends on subtle pixel differences, which, while effective for simple cognitive tasks, fall short in generalizing to the human internal understanding of the world.

Relative Embeddings: Training Anchors Similarity

Nevertheless, even with the inclusion of relational Vision Transformer embeddings, certain relevant anchor images continue to be retrieved. This demonstrates that relying solely on pixel-wise characteristics is valuable, yet insufficient. Through a Hadamard multiplication of KGE and ViT embeddings, we quantify the extent of agreement between ViT and KGE regarding an image's similarity to prototypes. As a result, KGE's semantics are maintained, but the images are re-ranked based on perceptual features detected by ViT. The results in Figure IV.3.8 underscore the promise of the hybrid approach, which leverages both relative representations. Notably, it excels in identifying semantically similar anchors, as illustrated in the cases of *fitness* and *danger*. This

superior performance in recognizing relevant similarities can be attributed to the complementary nature of relative ViT and relative KGE embeddings. When combined, they capture information that is sometimes missed when using them individually. These findings suggest that the hybrid approach offers significant benefits, particularly in situations where accuracy and semantic understanding are essential. By combining relative representations, it's possible to mitigate limitations present in either ViT or KGE embeddings, resulting in more robust and effective image classification. The hybrid approach shows promise for a wide range of applications where understanding the underlying factors contributing to image similarity is of paramount importance.

IV.3.5.4 Limitations and Further Directions

This section outlines several promising future directions that can significantly expand the horizons of research and enhance the effectiveness of our approach:

Better Anchors Our choice of anchors, involving the simple sampling of 100 images from each cluster, overlooks the complexity of image distribution. Implicitly assuming that all images within a cluster are identically and independently distributed (iid) is not reflective of reality, as our experiments on intraclass variance based on deep features in ARTstract have shown (see Chapter II.2). Instead, AC clusters contain a diverse range of samples, and the similarity between two images belonging to different clusters is likely higher than the similarity between images belonging to the same clusters. Although we sample a significant number of images (100) for each cluster to mitigate this issue, future work should focus on developing a more robust prototype selection strategy.

Pre-clustering for Semantic Anchors Introducing a pre-clustering step based on deep features, such as the Visual Transformer (ViT) features, is an intriguing concept. This step could involve identifying semantic clusters within each AC image, potentially based on visual features extracted by ViT. By pre-defining a set of semantically diverse clusters for each AC, one may be able to identify better anchors for relative representations. For instance, clusters associated with the *danger* category might include diverse clusters of images, some related to the ocean, others to war scenes, others to sharp objects, and more. This pre-clustering can enhance the granularity of semantic understanding and potentially improve classification accuracy.

Exploring Graph Neural Networks (GNNs) One promising direction for further research involves transitioning from traditional KG embeddings to Graph Neural Networks (GNNs) [239, 191]. This transition presents an opportunity to assess whether GNNs can outperform KG embeddings in the context of AC image classification. Furthermore, a fascinating aspect of this approach is the exploration of information flow within the GSNN. Understanding how information propagates through the network, including the involvement of specific nodes and edges, can provide valuable insights. One

approach to achieve this is by investigating the sensitivity of the GSNN's output to hidden states and activations, which can be accomplished by computing partial derivatives concerning the category of interest. However, while GNNs excel at capturing intricate relationships in graph structures, their advantage over KG embeddings depends on factors such as dataset characteristics and embedding techniques. Empirical validation through comparative studies is essential to determine the most effective approach for AC image classification.

Creating Scene Graphs To enhance the representation of perceptual semantics, another avenue for exploration is the construction of a scene graph [233, 152] for each image. Scene graphs offer the potential to capture intricate relationships between objects, actions, and attributes. This approach aligns with research in the field, and may uncover previously unnoticed relationships, adding a layer of complexity to the understanding of perceptual semantics in images.

Leveraging Multimodal Knowledge Fusion An intriguing avenue is to explore the fusion of perceptual knowledge with other modalities, such as textual information or audio data if available. This multimodal fusion can lead to a more comprehensive understanding of the content within images and open doors to various cross-modal tasks, including image captioning, audio-visual recognition, and more.

Knowledge-Driven Analysis of Misclassifications and Bias By harnessing the explicit knowledge graph, we can systematically collect and analyze misclassifications, leveraging SPARQL queries to identify patterns of errors (e.g., many images misclassified as *danger* when they should be categorized as *freedom* depict a certain object, like guns. The fact that we have an explicit KG enables this line of work, allowing us to explore and understand the root causes of misclassifications. In this way, we can scrutinize how biases may propagate through the KG and affect classification outcomes. Understanding these dynamics within the KG is essential for developing strategies to mitigate bias and ensure fairness in AC image classification.

Applying Self-supervised Learning Techniques Self-supervised learning approaches, where the model learns from unlabeled data, can be a promising direction for further refining either anchors of the KG embeddings.

IV.3.6 Conclusions

This chapter introduced the AKG, a versatile and context-aware knowledge repository capturing the perceptual semantic attributes of over 14,000 cultural images tagged with ACs. The AKG reifies perceptual semantics and explicitly encodes when, where, and how these annotations were made, along with annotation strength, effectively rendering these annotations *situated* ground truths. It also establishes connections with commonsense knowledge sources like ConceptNet, enabling interpretable reasoning over perceptual semantic data. The AKG thus serves as a foundational resource for contextual understanding in visual sense-making, laying the groundwork for further research and analysis.

Furthermore, we have showcased the potential of Knowledge Graph Embeddings in enhancing AC image classification accuracy and their complementarity when combined with Vision Transformer embeddings. We exploited the AKG for the task of AC image classification through the use of KGE–both in their absolute and relative representations. Our results revealed the significant potential of the relative representation method to boost the performance of KGE-based models. In addition, our experiments explored hybrid approaches that combined KGE and ViT embeddings to assess their joint utility for AC image classification. These hybrid embeddings demonstrated their ability to enhance classification accuracy significantly, emphasizing their complementarity in capturing image relationships and shared attributes. Our results surpassed the state of the art in AC image classification.

Additionally, our interpretability experiments have provided valuable insights into the models' behaviors, underlining their unique strengths and weaknesses. While ViT excelled at capturing detailed visual features, KGE demonstrated a superior capacity for interpreting scenes, high-level semantics, and complex interactions in images. This discrepancy could be attributed to the prototype selection process and the inherent noise within ViT's latent space. The relative representation approach introduced a prior assumption, confining the representation to specific regions within the latent space. This constraint could limit ViT's ability to express semantic relationships and result in perplexing images that differ significantly from human perception. These findings raise critical questions about the interpretability and semantic understanding of images depending on the models learning their representations.

Our findings, while promising, also reveal the challenges and nuances in interpreting and understanding perceptual semantics within images. We have identified several avenues for future research, including the exploration of better anchor selection strategies, the integration of Graph Neural Networks (GNNs), the construction of scene graphs, and the utilization of multimodal knowledge fusion. We also emphasized the importance of knowledge-driven analysis of misclassifications and bias, as well as the potential of self-supervised learning techniques to further refine image representations.

Part V

Conclusions

Chapter V.1

Towards Hybrid Cognitive AI

V.1.1 Summary of Research Objectives

This dissertation tackled the challenging task of bridging the semantic and cultural gaps between raw pixel data and high-level visual semantics, with a focus on AC image classification. Its goal was to improve the performance and interpretability of current state-of-the-art approaches. The central hypotheses explored the potential of cognitive-inspired, automatically detectable semantic intermediaries as proxies for AC evocation, highlighting the role of semantic technologies in developing efficient and interpretable hybrid intelligent systems.

To support this research, we introduced the ARTstract image dataset (Chapter II.1). This dataset comprises 14,000 cultural images, including artworks, historical photographs, and advertisements, carefully curated from four image datasets [2, 338, 390], and providing a unique resource for investigating the intersection of visual data and ACs. Apart from using it as the foundation for our experiments, ARTstract serves as a solid starting point for further research in areas such as CV, digital art history, and cognitive science, enabling experimentation with interpretable CV methods and hopefully inspiring the development of more culturally sensitive and diverse datasets for high-level visual semantics.

We established our technical research within the domain of multi-class image classification. Given the ARTstract dataset comprising images $X = [I_1, I_2, \ldots, I_m]$, each paired with corresponding ground truth labels $Y = [y_1, y_2, \ldots, y_m]$, with each label being selected from a set of K potential AC classes. Our research task was to ascertain the optimal image representation, I_i, and model parameters, θ, that enable us to predict the label $\hat{y_i}$ in a way that it closely aligns with the true label y_i. We expressed this objective using the following equation:

$$\hat{y_i} = \arg\max(p(y_i|I_i, \theta))$$

Our research tackled this challenge by exploring how various machine-readable image representations would influence the performance and interpretability of an AC image classification system. We aimed to uncover insights into the most informative and interpretable aspects of image data by applying transformation functions to the original raw pixel representation, I_{RAW}. Our investigation revolved around four distinct paradigms, each with unique attributes (see Table V.1.1).

Insights from the Background Research

To lay the foundations for the proposed methods, in Chapter I.2, we systematically explored the landscape of high-level visual understanding, focusing on identifying work that explicitly or implicitly dealt with the task of AC image classification. Through a multidisciplinary approach, we defined and characterized "high-level semantics" by identifying the semantic units assigned to this tier. We then classified these semantic units across four knowledge clusters: commonsense, emotional, aesthetic, and interpretative semantics. This approach allowed us to then survey and identify past CV work for associating these high-level semantic units with images. Our survey identified a substantial body of work and categorized the identified CV tasks into five analogous clusters: situational analysis, visual sentiment analysis, aesthetic analysis, social signal processing, and visual rhetorical analysis.

Our findings revealed that the field of CV is increasingly shifting its focus towards the automatic detection of sociocultural and subjective elements within images, including emotions, personality traits, and visual rhetorics. Furthermore, the survey showcased a strong reliance on Convolutional Neural Networks (CNN) and an expansion of tasks from natural photographs to cultural images, highlighting the significance of tailored datasets and data augmentation. Most significantly, the survey underscored the daunting challenge of achieving high F1 scores in AC image classification, even with substantial datasets, prompting a reevaluation of the data accumulation paradigm. The integration of symbolic knowledge and the recognition of mid-level features, such as objects and actions, emerged as pivotal strategies to enhance performance in this intricate task.

Furthermore, Chapter I.3 set the stage by providing a succinct foundation on the cognitive science perspective on ACs and their representation within the human brain. We highlighted the coexistence of distributional and embodied information in grounding ACs and examined key cognitive aspects, including acquired embodiment, relationality, and emotionality. This foundational knowledge served as a critical underpinning for the practical applications discussed in the subsequent chapters of AC image classification.

Paradigm	Approach	Image Representation	Task Formulation
End-to-end Deep Vision (DL)	Pre-trained DL Models + Finetuning	$f_{DL} : I_{RAW} \rightarrow I_{DL} \subseteq \mathbb{R}^{768}$	$\hat{y} = \arg\max(p(y_i \mid I_{\text{DL}}, \theta))$
Perceptual Semantics (PS)	Label Extraction + Feature Engineering + Classical ML	$f_{PS} : I_{RAW} \rightarrow I_{PS} \subseteq \mathbb{R}^{N}$	$\hat{y} = \arg\max(p(y_i \mid I_{\text{PS}}, \theta))$
Situated Knowledge of Perceptual Semantics (SKPS)	Perceptual Semantics + Annotation Situations + Commonsense Frames + KG Embeddings	$f_{KG} : I_{PS} \rightarrow I_{KG} \subseteq \mathbb{G}$ $f_{KGE} : I_{KG} \rightarrow I_{KGE} \subseteq \mathbb{R}^{128}$	$\hat{y} = \arg\max(p(y_i \mid I_{\text{KGE}}, \theta))$
Hybrid: End-to-end Vision + Situated Perceptual Knowledge	Concatenation of Absolute Embeddings	$f_H : [I_{DL}; I_{KGE}] \rightarrow I_H \subseteq \mathbb{R}^{896}$	
	Concatenation of Relative Embeddings	$f_H : [I_{R-DL}; I_{R-KGE}] \rightarrow I_H \subseteq \mathbb{R}^{1400}$	$\hat{y} = \arg\max(p(y_i \mid I_{\text{H}}, \theta))$
	Hadamard Product of Relative Embeddings	$f_H : I_{R-DL} \odot I_{R-KGE} \rightarrow I_H \subseteq \mathbb{R}^{700}$	

Table V.1.1: Comparative overview of the four paradigms investigated in this dissertation. Each paradigm offers a unique approach to transforming image representations for use in the AC image classification task. The explored paradigms include end-to-end deep vision, perceptual semantics with classical machine learning, situated knowledge of perceptual semantics with KG embeddings, and a hybrid approach.

V.1.2 Addressing our Research Questions

This dissertation research was structured around three goals, each accompanied by one of the three central research questions (RQs) (see introductory Section I.1.4) guiding this work. In this section, we discuss how we answered those RQs.

> **1. Deep Learning for Abstract Concept Image Classification**: Investigating the effectiveness of state-of-the-art DL models in handling ACs in image data through an end-to-end approach.
>
> **RQ1**. *To what extent can the end-to-end DL paradigm, connecting raw pixel values directly to unsituated AC labels, address the task of AC image classification in terms of both performance and explainability?*

We explored this research question in Chapter II.2, conducting a critical assessment of the end-to-end DL paradigm's performance, representation, and interpretability. Our findings reveal that this paradigm presents significant challenges when applied to AC image classification. Specifically, we utilized state-of-the-art pretrained DL models (VGG-16, ResNet-50, and ViT) to transform raw pixel-based images (I_{RAW}) into deep feature vectors (I_{DL}) with dimensions determined by the reused model. This process involved fine-tuning the pre-trained DL models to enable direct image classification using these deep features. In our analysis of VGG-16 image feature vectors (Section II.2.4), we identified significant differences in intraclass similarity between ARTstract and CIFAR-10 classes, suggesting a correlation between target class abstractness and the model's ability to capture shared features. Concrete classes demonstrated high intraclass similarity, signifying effective feature extraction by the CNN-based model. Conversely, AC classes showed lower similarity, highlighting the challenge of AC-relevant feature extraction from images with DL models in an end-to-end approach. In the finetuning and performance evaluation of deep models (Section II.2.5), we found that DL models excelled in standard image classification, but struggled with ACs. The notable contrast in accuracy between CIFAR-10 (concrete classes) and ARTstract (abstract classes) underlines the difficulties of dealing with ACs in visual content and the limitations of conventional DL models. These results emphasize the complexity of AC classification, characterized by open definitions, polysemy, and diverse associations with visual data. However, our ARTstract-trained models achieved higher F1 scores than those trained on other datasets, highlighting the potential of the ARTstract dataset in advancing AC image classification. Finally, in Section II.2.6, we explored various methods to interpret the knowledge acquired by the fine-tuned models. Conventional explainability techniques like Grad-CAM had limitations in revealing the intricate relationship between visual data and ACs due to their reliance on concrete evidence. The challenges posed by high intraclass dissimilarity in visual representations of ACs impacted both model performance and the clarity of traditional feature visualizations. Our innovative neuron visualization denoising approach, SD-AM, proved partly effective in generating human-readable "hypericons" that capture features DL model associated with target ACs. Interestingly,

we observed a visual convergence observed between modern hyperpop aesthetics and SD-AM hypericons. This prompts contemplation about the relationship between contemporary media consumption and the data processing capabilities of DL models, as shown in Figure II.2.18. In conclusion, this chapter made a substantial contribution to our comprehension of the strengths and weaknesses of the end-to-end DL paradigm in the context of AC image classification.

> **2. Minding the Gap with Cognitive Intermediaries**:
> Exploring the potential of visual data descriptors to bridge the gap between raw pixels and ACs via perceptual semantics.
>
> **RQ2**. *Is it possible to automatically identify intermediary semantic features to bridge the semantic gap between raw pixels and ACs? How might the utilization of these features impact the performance and interpretability of AC image classification?*

This research question was addressed in Part III, which introduced the *perceptual semantics* (PS) paradigm to bridge the semantic gap between raw pixels and ACs in the context of image classification. The PS paradigm aimed to close this gap by automatically extracting concrete labels (such as objects, actions, emotions, colors, and more) from images, and using these semantic labels as explicit semantic intermediaries. Unlike end-to-end DL approaches, the PS paradigm follows a feature engineering strategy, converting raw images (I_{RAW}) into perceptual semantic representations (I_{PS}) that explicitly correspond to the presence of concrete perceptual semantics. As a result, the I_{PS} representation is characterized by a more interpretable, symbolic foundation.

Specifically, in Chapter III.1, we presented empirical evidence that showcases the potential of computational techniques to capture symbol grounding and *acquired embodiment*. This evidence was derived from a case study using visual artworks from the Tate Gallery. The study identified commonsense perceptual features that frequently co-occur with ACs. Notably, associations were found between AC tags, like *consumerism*, and tag descriptors such as "clothing," "food and drink," and "product packaging." Additionally, our color analyses provided insights into the chromatic attributes of AC visual evocation within the Tate collection.

Chapter III.2 extended the acquired embodiment software architecture, with a focus on improving the interpretability of image representations. This chapter explored the use of DL models at more concrete layers of the semantic pyramid. Perceptual semantics were autonomously extracted for all ARTstract images, leading to more interpretable image representations. We then harnessed these image representations to train interpretable classical Machine Learning (ML) techniques, including Naive Bayes, to perform AC image classification. The results demonstrated the effectiveness of this approach, which maintained performance levels similar to convolutional neural networks (CNNs) while significantly enhancing interpretability. This exploration challenged the prevailing reliance on end-to-end DL for abstract and subjective tasks, highlighting the potential of keeping these methods at shallower levels of the semantic pyramid. Overall, our results underscored the effectiveness of feature engineering and traditional ML techniques,

emphasizing their significance in socio-cultural cognitive tasks where interpretability is essential.

3. Reifying and Reasoning with Knowledge Graphs:
Analyzing the potential of semantic technologies in representing the commonsense and cultural dimensions of perceptual semantics to enhance AC image classification.

RQ3. *How can the fusion of perceptual semantics with symbolic knowledge through ontology-based frameworks enhance the performance and interpretability of AC image classification?*

In Part IV we provided an answer to this research question by introducing the *situated perceptual knowledge* (SPK) paradigm. This paradigm extended the principles of the PS paradigm, aiming to bridge the semantic gap, tackle commonsense reasoning, and address the cultural gap. It utilized ontology-based KGs to reify semantic relationships, facilitating reasoning and integration with common sense knowledge. The transformation of perceptual semantic representations of images (I_{PS}) into structured KG format (I_{KG}) was a core aspect. This KG was further embedded into a vector space to create a vector representation (I_{KGE}) used for AC image classification. The paradigm's foundation rests on two critical contributions: a process for reasoning over linguistic visual descriptors through the incorporation of commonsense knowledge (Chapter IV.1) and the introduction of the SituAnnotate ontology-based framework for semantically situating this knowledge (Chapter IV.2). Chapter IV.3 elaborated on the application of these contributions within the context of the ARTstract dataset for AC image classification, providing insights and solutions to the research question.

Specifically, Chapter IV.1 explored the connections between perceptual semantics descriptors and linguistic frames with an emphasis on interpretability. Our findings confirmed that the use of concrete textual label descriptors can establish meaningful links between image perceptual semantics and higher-level concepts while maintaining interpretability through commonsense reasoning. This approach, incorporating ontology-based knowledge engineering techniques and commonsense knowledge, not only enhanced visual material descriptions but also simplified the retrieval of visual instances of semantic frames, promoting two-way information exchange. This research signifies a notable advancement, demonstrating the potential of such methodologies to bridge the gap between specific image labels and high-level frames.

Subsequently, Chapter IV.2 explored the presence of subjective and cultural biases solidified in the extraction of PS, and introduced methods to address these biases. Our results highlighted the challenges in achieving objectivity in data annotations and the impact of cultural perspectives on image labeling, including in gold standard CV datasets like Visual Genome [233]. We introduced SituAnnotate as an alternative to traditional annotation methods, to improve annotation precision and contextual depth. Its adaptable architecture accommodates various annotator types while ensuring contextual grounding and aligning with modern annotation practices. Furthermore,

our results showed that automated reasoning through SPARQL queries streamlined annotation comprehension, offering nuanced insights into annotation contexts. By contextualizing annotations, SituAnnotate effectively mitigates bias and promotes ethical AI development, meeting the demand for transparent, responsible, and bias-aware AI systems. Our practical demonstration underlined SituAnnotate's value in providing insightful and human-understandable explanations for annotations, enhancing contextual understanding, and improving the quality of AI development practices.

Finally, in Chapter IV.3, we explored how grounding perceptual features in their subjective, cultural, and commonsense contexts can enhance AC image classification performance and interpretability. We achieved this by creating the ARTstract Knowledge Graph (AKG), an expansion of the ARTstract image dataset. We integrated perceptual semantic features with situational metadata and high-level linguistic frames, creating multifaceted image representations within the KG. This representation provided a nuanced understanding of images by considering the origins, strengths, and interconnections of descriptors. We employed the TransE embedding method to obtain Knowledge Graph Embeddings (KGE) for image representation, and conducted multiple experiments for AC image classification using KGEs exclusively. Our findings demonstrated that the KGE approach, while not reaching the Vision Transformer (ViT)'s level of performance, experienced significant performance improvement when incorporating relative representations [261]. This superior performance indicates that this method introduces a more 'semantic' representation, thereby boosting semantic resolution. In a broader context, these findings emphasize the potential of the relative representation method to enhance KGE-based image classification, offering a valuable alternative to ViT.

Furthermore, through our hybrid approaches that combine deep embodied features with symbolic representations, we investigated multiple fusion strategies. Some of these strategies incorporated relative representations to establish effective connections between KGE and ViT, all while ensuring invariance to latent isometries and rescalings. Our interpretability analyses shed light on the distinct capabilities of ViT and KGE in comprehending high-level semantic concepts. ViT primarily focused on lower-level visual features such as colors and textures, whereas in some cases KGE excelled at capturing more abstract and high-level scene aspects (e.g., Figure IV.3.7). Importantly, our hybrid methods lead to improvements in AC image classification performance, surpassing all available state-of-the-art methods. This chapter demonstrated multiple ways in which ontology-based frameworks can enhance image classification in terms of both performance and interpretability, significantly contributing to our response to RQ3. These findings also underscored the potential of neuro-symbolic methods, including KGs and their embeddings, for robust image representation in complex visual understanding tasks, confirming their value in CV and AI research.

V.1.3 Key Research Contributions

In this section, we provide a concise synthesis of our key research contributions:

V.1.3.1 End-to-end DL Vision

- Introduction of *ARTstract* image dataset, only dataset we know to explicitly focus on AC tags, containing +14K cultural images, including advertisements and artworks, and offering the potential for enhancing AC image classification performance and future work.
- Benchmark performances on ARTstract using state-of-the-art DL models.
- Identification of the strengths and challenges posed by the *end-to-end DL paradigm* for AC image classification.
- Introduction of innovative interpretability techniques like SD-AM *hypericons* for gaining insights into the cultural meanings learned by DL models.

V.1.3.2 Minding the Gap with Cognitive Intermediaries

- Empirical evidence of the feasibility of translating the cognitive theory of ACs' *acquired embodiment* into a computer vision-based software approach.
- Introduction of the *perceptual semantics* paradigm, which uses DL at more concrete layers of the semantic pyramid, extracting intermediary semantic features to bridge the semantic gap between raw pixels and ACs.
- Evidence of the effectiveness of combining feature engineering and classical ML techniques to improve interpretability while maintaining performance levels comparable to CNNs.
- Challenge to the prevailing reliance on end-to-end DL for abstract and subjective CV tasks, highlighting the potential of alternative approaches.

V.1.3.3 Reifying and Reasoning with KGs

- Introduction of an interpretable commonsense reasoning approach to automatically connect visual data with high-level linguistic frames.
- Introduction of the *SituAnnotate* ontology-based framework, enhancing annotation precision and depth, and addressing subjective and cultural biases in data labeling for improved fairness and accuracy.
- Introduction of the *ARTstract Knowledge Graph*, expanding the ARTstract image dataset by situating perceptual semantics annotations for bias awareness and enhancing symbolic-embodied knowledge.
- Introduction of the *situated perceptual knowledge* (SPK) paradigm, which extends the PS paradigm by reifying semantic relationships and using ontology-based KGs to enhance AC image classification.

- Introduction of an AC image classification method outperforming all state-of-the-art methods and even surpassing the top end-to-end DL model (ViT).
- Quantitative and qualitative evidence supporting the value of neuro-symbolic methods, including KGs and their embeddings, for robust image representation in high-level visual sensemaking tasks, with implications for performance and interpretability.

V.1.4 Open Questions and Future Directions

The research in this thesis highlights the nuanced balance between simplifying complexity for manageability and enriching complexity for deeper comprehension in high-level visual understanding and AC image classification. A central theme throughout the dissertation is navigating this balance using intermediaries to bridge the semantic gap between raw pixel data and ACs. While these intermediaries aid in explainability and reasoning, it is essential to acknowledge the potential drawbacks of early complexity reduction. For instance, simplifying complexity through intermediaries, such as identifying dominant colors or other low-level features, may oversimplify the nuanced information within visual data. This oversimplification could lead to overlooking crucial subtleties and variations necessary for accurate interpretation. While reductionist approaches enable more manageable analysis, it's imperative to exercise caution to prevent limiting outcomes to specific conclusions. To address this challenge, potential strategies include integrating hierarchical structures to iteratively refine interpretations and devising mechanisms to adjust abstraction levels dynamically based on contextual cues.

As we conclude this work, it is essential to identify promising directions for future research and development in these areas. The following points outline potential avenues for further exploration:

V.1.4.1 User Study for AC Image Classification

We have primarily evaluated the performance of our AC image classification methods using quantitative measures such as accuracy and F1 scores. While quantitative metrics provide valuable insights into the performance of AI models, it is equally important to assess their real-world utility. A task-based user study could be designed to simulate scenarios where humans are required to classify images based on ACs, similar to the tasks that our AI models perform. This user study should involve participants from various backgrounds, including those unfamiliar with the domain. Key objectives of such a study would include:

- Assessing how well humans perform on AC image classification tasks, providing a benchmark against which AI model performance can be compared.
- Evaluating the consistency and subjectivity of human judgments when identifying ACs in images, shedding light on the inherent challenges of these tasks.

- Exploring areas where AI models outperform or underperform humans and vice versa, contributing to a more nuanced understanding of their respective strengths and limitations.
- Gathering qualitative feedback from users about their experiences, difficulties, and insights while performing AC image classification, which can inform the development of more cognitive-inspired AI systems.

V.1.4.2 User Evaluation for Explainability

Understanding how well users can comprehend the rationale behind an AI model's decisions is vital for building trust and transparency in AI systems. Future research should consider conducting user studies to assess the explainability of our and other methods. This can involve presenting users with AI-generated explanations for specific image classifications and gathering their feedback on the clarity, comprehensibility, and utility of these explanations. Users' perceptions of which methods are most effective in helping them understand how an AI model reached its conclusions can guide improvements in explainability techniques. Incorporating user feedback and preferences can lead to the development of more user-centric AI systems and further enhance their utility in real-world applications.

V.1.4.3 Dataset Expansion and Diversity

While ARTstract serves as a pioneering dataset for AC image classification, further expansion and diversity can enhance its potential. Future work could focus on enriching the dataset with a more extensive range of cultural images, including those from non-Western perspectives, to capture a broader spectrum of ACs. This increased diversity could better represent the cultural richness and subjectivity inherent in high-level visual understanding. To achieve this, the following steps can be considered:

- Obtain human annotations of ACs to assess how ARTstract's annotations compare to those provided by annotators from different cultural backgrounds. Incorporating human-checked tags and tracking them using SituAnnotate would significantly enhance ARTstract's reliability and power as a resource.
- Expand the cluster definitions within ARTstract, encompassing a wider array of words, languages, and cultural nuances. This expansion would make ARTstract more comprehensive and inclusive, improving its utility in AC image classification across diverse cultural contexts.

V.1.4.4 Refining Task Definitions and Evaluation Metrics

Refining the task definition and evaluation metrics for AC image classification is a critical avenue for future research. The conventional multi-class classification approach may not

be the most suitable, given that many ACs can be associated with a single image. To better align the task with human vision and cognition, alternative task definitions and evaluation metrics should be explored. Here are some proposed directions:

- Adopting a ranking-based task rather than traditional classification, which can provide a more nuanced perspective on the relative relevance of different ACs to an image. We could follow an approach similar to [187], which proposes a relative scale for ambiguous labels and compares images as a ranking task.
- Adopting a multi-label multi-class classification paradigm instead of the single-label multi-class paradigm that has been adopted in this thesis. This shift acknowledges that an image may be associated with multiple ACs simultaneously, allowing for a more comprehensive understanding of the content's nuances and complexities.
- Prioritizing reasonability over objectivity for evaluation, as proposed by [2]: This method seeks reasonable associations instead of ground truths, asking evaluators to rate the reasonableness of annotations, embracing subjectivity and finding common ground.
- Developing evaluation metrics that take into account the semantic relationships between different AC classes: For instance, if an image is initially labeled as *death* but is mistakenly classified as *danger*, the evaluation metric may assign a distinct score compared to when it is misclassified as *comfort*. Evaluation metrics could explicitly address these nuances, potentially using multi-label learning with varying label importance and predicting AC probability distribution, as suggested in the context of visual sentiment analysis for emotions [405].

V.1.4.5 Expansion to Natural Images

A logical next step is to broaden the scope of our methods to encompass natural images, incorporating a wider array of images and datasets. This extension will facilitate the evaluation of AC image classification techniques in a broader context, one not confined to cultural or artistic images. By exploring natural images, we can gauge how well our methods generalize to everyday visual content, thus achieving a more comprehensive understanding of their effectiveness across various domains. Furthermore, directing attention to natural images may unveil distinct challenges and opportunities unique to this context, contributing to the enrichment of the research landscape.

V.1.4.6 Ethical Implications, Bias, and Fairness

To investigate the ethical implications of AC image classification systems, it is essential to better understand their potential for misuse in surveillance, governance, propaganda, and other contexts. Research in this direction can shed light on the risks associated with these systems and guide the development of ethical guidelines and safeguards.

Future research should focus on developing robust methods for detecting, mitigating, and ensuring transparency in both cultural and subjective biases. This encompasses strategies for creating and maintaining fair and ethical datasets, annotation processes, and AI systems. Emphasizing bias-aware AI development will be vital in tackling the challenges posed by the cultural and subjective interpretations of visual content.

V.1.4.7 LLMs

While LLMs (Large Language Models) have primarily been associated with natural language understanding and generation, there is an emerging potential to leverage these models for high-level visual understanding tasks, such as AC image classification. Future research in this area should explore the integration of LLMs with existing DL and CV techniques to enhance the performance and interpretability of AC image classification. Additionally, investigating how LLMs can assist in refining semantic representations and improving the alignment between textual descriptions and visual content in the context of ACs could be a promising direction. In Section V.1.5.3 of the Appendix, we present some initial results and further discussion.

V.1.5 (Taming) Wicked Problems

The challenges encountered in this dissertation highlight the intricate, queer, and dynamic nature of high-level visual understanding and ACs. As we conclude this work, we advocate for an approach that *not only acknowledges but embraces* the intricate and multifaceted characteristics of ACs in an era of algorithmic curation. We propose using the concept of 'wicked problems' [297] as a lens to better comprehend and address these complexities, emphasizing that embracing this complexity is paramount for the ethical development of AI.

V.1.5.1 Embracing the Queer Complexity of ACs

The results of this dissertation highlight the "inherently queer" and transgressive nature of these concepts [247] and their ability to transcend the limitations of binary confinement. In this context, the term "queer" encompasses a discourse that challenges conventional binary cultural paradigms and suggests a more fluid understanding [7]. This queerness implies a recognition of the incompleteness of definitions and invites novel perspectives, as elaborated by Butler [63, p. 228],

> If the term "queer" is to be a site of collective contestation, the point of departure for a set of historical reflections and futural imaginings, it will have to remain that which is, in the present, never fully owned, but always and only redeployed, twisted, queered from a prior usage and in the direction of urgent and expanding political purposes, and perhaps also yielded in favor of terms that do that political work more effectively.

Importantly, the queerness of ACs arises from their resistance to fixed, simplistic, or normative definitions, a resistance that poses both significant technical and ethical challenges due to their elusive, subjective, and context-dependent nature. However, understanding the challenges posed by the inherently queer nature of ACs allows us to approach their automatic labeling and detection with a broader awareness. This awareness aligns with Butler's assertion that the term "queer" remains in a state of constant transformation, evolving in response to ever-expanding political contexts [63]. The idea of perpetual redeployment and twisting applies aptly to the dynamic landscape of ACs, where their interpretation and application continue to evolve, shaped by diverse perspectives and emerging contexts.

V.1.5.2 Leveraging Wicked Problem Solving

The concept of ongoing transformation aligns with 'wicked problems' [297], a framework rooted in social policy planning that acknowledges the multifaceted and evolving nature of complex 'wicked' challenges compared to 'tame' ones. Tame problems have well-defined boundaries and a limited number of factors, leading to unequivocal right-or-wrong solutions. Wicked problems, on the other hand, exist within open systems encompassing multidimensional, multicultural, and robust cultural dimensions, making them socially intricate and open-ended. Unlike tame problems with clear solvability, wicked problems lack a definitive test, resulting in assessments on a continuum of better or worse outcomes. Addressing wicked problems necessitates platforms for a shared understanding of complexity, a paradigm integrating various contexts, and analyses of power dynamics. An iterative, experimental approach is essential, recognizing the evolutionary nature of problems and solutions. Thus, in contrast to tame problems with clear-cut solutions, wicked problems like automatic AC image classification require a paradigm shift toward collaborative thinking and diverse perspective integration to tackle their complexity.

I propose that computer science and CV can gain valuable insights by adopting a wicked problem perspective in the realm of high-level visual sensemaking. This approach involves recognizing and embracing specific attributes:

- **Openness**: Automatic high-level visual understanding tasks, like image classification based on ACs, occur within open systems that transcend spatial and temporal constraints. For instance, interpreting an AC like *freedom* in images involves understanding it across different cultures and historical contexts, requiring sensitivity to diverse meanings.

- **Multidimensionality**: Image classification involving ACs requires understanding images from various angles—cultural, social, and human values. Consider the previously discussed *Tank Man* image from Tiananmen Square. The multiple, even antonymic interpretations illustrate the multi-layered nature of the evocation of ACs from visual data.

- **Evolving nature**: The problem-solving process in automatic high-level visual understanding has to evolve as concepts and contexts change. For example, as societal perceptions of *gender* evolve, the way images depicting gender-related concepts are classified must also adapt to reflect current understandings.
- **Complexity**: Complexity in automatic high-level visual understanding arises from the interplay of social and cultural factors shaping image interpretations. Beyond visual cues, understanding cultural nuances, historical contexts, and societal dynamics is vital. For instance, consider a courtroom image; to some, it might symbolize *justice*, but for others, it could represent *systemic biases*.
- **Subjectivity**: Automatic high-level visual understanding tasks inherently involve subjectivity. For instance, interpreting whether an image evokes *success* depends on individual perspectives, influenced by cultural norms, personal values, and societal expectations, making consensus challenging.
- **Power dynamics**: The issue of power dynamics in automatic high-level visual understanding is evident when classifying images depicting ACs like *authority* or *purity*. Different stakeholders may interpret the image differently based on their positions and backgrounds, highlighting the need to address power imbalances in labeling processes.

Embracing a set of such fundamental values demands not only technical advancements but also the integration of interdisciplinary methodologies.

V.1.5.3 Concluding Thoughts

As we embark on the era of algorithmic curation, it becomes increasingly apparent that the intricate interplay of visual data and cultural connotations cannot be solely governed by technical performance metrics. The quest for machine intelligence to interpret and categorize ACs needs a convergence of technical innovation with an awareness of social contexts. This dissertation explored and probed the nuanced dynamics intrinsic to ACs within the landscape of CV, emphasizing the challenges resulting from the semantic and cultural gaps. Yet, it is within these gaps that the true richness of meaning is found, reflecting the dynamic and diverse nature of human perception and cognition. I finish by spotlighting the limitations inherent to binary thinking, proposing to address these intricate issues through the lens of 'wicked problems,' rich with complexity and multidimensionality. By embracing a vantage point that encompasses a situated understanding of ACs, we forge a path toward responsible labeling and training over visual media. The insights gleaned from this work emphasize the vital role of interdisciplinary collaboration, fostering a culture of critical inquiry, and cultivating a queer perspective in reshaping the horizons of CV and its profound societal impact. The synthesis of contemporary cognitive neuroscience, ethical sensibilities, and technical innovation propels us to usher in an era of AI development that harmonizes technical prowess with a profound awareness of the human and societal dimensions it touches upon.

List of Figures

Bibliography

[1] Yalemisew Abgaz et al. "A Methodology for Semantic Enrichment of Cultural Heritage Images Using Artificial Intelligence Technologies." en. In: *Journal of Imaging* 7.8 (2021), p. 121. DOI: 10.3390/jimaging7080121.

[2] Panos Achlioptas et al. "ArtEmis: Affective Language for Visual Art." en. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, TN, USA: Computer Vision Foundation / IEEE, 2021, pp. 11569–11579. DOI: 10.1109/cvpr46437.2021.01140.

[3] Somak Aditya, Yezhou Yang, and Chitta Baral. "Explicit reasoning over end-to-end neural architectures for visual question answering." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 2018.

[4] Somak Aditya, Yezhou Yang, and Chitta Baral. "Integrating knowledge and reasoning in image understanding." In: *28th International Joint Conference on Artificial Intelligence, IJCAI 2019*. International Joint Conferences on Artificial Intelligence. 2019, pp. 6252–6259.

[5] Imad Afyouni, Zaher Al Aghbari, and Reshma Abdul Razack. "Multi-feature, multi-modal, and multi-source social event detection: A comprehensive survey." In: *Information Fusion* 79 (2022), pp. 279–308.

[6] Siddharth Agarwal et al. "Genre and style based painting classification." In: *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE. 2015, pp. 588–594.

[7] Sara Ahmed. *What's the use?: On the uses of use*. Duke University Press, 2019.

[8] Youssef Ahres and Nikolaus Volk. "Abstract Concept and Emotion Detection in Tagged Images with CNNs." en. In: *Unpublished Report, accessed from http://cs231n. stanford. edu/reports/2016/pdfs/008_ Report. pdf* (2016), p. 8.

[9] Naveed Akhtar et al. "Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey." In: *IEEE Access* 9 (2021), pp. 155161–155196. DOI: 10.1109/access.2021.3127960.

[10] Hala Al Kuwatly, Maximilian Wich, and Georg Groh. "Identifying and measuring annotator bias based on annotators' demographic characteristics." In: *Proceedings of the fourth workshop on online abuse and harms*. 2020, pp. 184–190.

[11] Jeanette Altarriba and Lisa Bauer. "The Distinctiveness of Emotion Concepts: A Comparison between Emotion, Abstract, and Concrete Words." In: *The American journal of psychology* 117 (2004), pp. 389–410. DOI: 10.2307/4149007.

[12] Jeanette Altarriba, Lisa M. Bauer, and Claudia Benvenuto. "Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words." en. In: *Behavior Research Methods, Instruments, &Computers* 31.4 (1999), pp. 578–602. ISSN: 1532-5970. DOI: 10.3758/bf03200738.

[13] Seyed Ali Amirshahi et al. "Jenaesthetics subjective dataset: analyzing paintings by subjective scores." In: *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13*. Springer. 2015, pp. 3–19.

[14] Mark Andrews, Stefan Frank, and Gabriella Vigliocco. "Reconciling Embodied and Distributional Accounts of Meaning in Language." en. In: *Topics in Cognitive Science* 6.3 (2014), pp. 359–370. ISSN: 1756-8765. DOI: 10.1111/tops.12096.

[15] Mark Andrews, Gabriella Vigliocco, and David Vinson. "Integrating experiential and distributional data to learn semantic representations." In: *Psychological Review* (2009), pp. 463–498.

[16] Matthew Arnold et al. "FactSheets: Increasing trust in AI services through supplier's declarations of conformity." In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 6–1.

[17] Taylor Arnold and Lauren Tilton. "Distant viewing: analyzing large visual corpora." In: *Digital Scholarship in the Humanities* 34.Supplement_1 (2019), pp. i3–i16. ISSN: 2055-7671. DOI: 10.1093/llc/fqz013.

[18] Taylor Arnold and Lauren Tilton. "Distant Viewing Toolkit: A Python Package for the Analysis of Visual Culture." In: *Journal of Open Source Software* 5.45 (2020), p. 1800. ISSN: 2475-9066. DOI: 10.21105/joss.01800.

[19] Lora Aroyo and Chris Welty. "Truth is a lie: Crowd truth and the seven myths of human annotation." In: *AI Magazine* 36.1 (2015), pp. 15–24.

[20] Sören Auer et al. "Dbpedia: A nucleus for a web of open data." In: *International Semantic Web Conference 2007*. Springer, 2007, pp. 722–735.

[21] Babajide O Ayinde, Tamer Inanc, and Jacek M Zurada. "On correlation of features extracted by deep neural networks." In: *2019 International Joint Conference on Neural Networks (IJCNN)*. Ieee. 2019, pp. 1–8.

[22] Reshmashree B Kantharaju et al. "Multimodal Analysis of Cohesion in Multi-party Interactions." In: *Lrec*. Marseille, France, 2020, pp. 498–507.

[23] Franz Baader et al. *The description logic handbook: Theory, implementation and applications*. Cambridge university press, 2003.

[24] C Babbage. *Passages From the Life of a Philosopher, ch. VIII*. 1864.

[25] Randheer Bagi, Tanima Dutta, and Hari Prabhat Gupta. "Deep learning architectures for computer vision applications: a study." In: *Advances in data and information sciences*. Springer, 2020, pp. 601–612.

[26] Matteo Baldoni et al. "From tags to emotions: Ontology-driven sentiment analysis in the social semantic web." In: *Intelligenza Artificiale* 6.1 (2012), pp. 41–54. DOI: 10.3233/ia-2012-0028.

[27] Hichem Bannour and Céline Hudelot. "Towards ontologies for image interpretation and annotation." In: *2011 9th International Workshop on content-based multimedia indexing (CBMI)*. Ieee. 2011, pp. 211–216.

[28] Iain Barclay et al. "Towards traceability in data ecosystems using a bill of materials model." In: *International Workshop on Science Gateways*. CEUR-WS. 2019.

[29] Lawrence W Barsalou. "Abstraction in perceptual symbol systems." In: *Philosophical Trans. of the Royal Society B: Biological Sciences* 358.1435 (2003), pp. 1177–1187. ISSN: 0962-8436.

[30] Lawrence W. Barsalou. "Perceptual symbol systems." en. In: *Behavioral and Brain Sciences* 22.4 (1999), pp. 577–660. ISSN: 0140-525x, 1469-1825. DOI: 10.1017/s0140525x99002149.

[31] Lawrence W Barsalou and Katja Wiemer-Hastings. "Situating abstract concepts." In: *Grounding cognition: The role of perception and action in memory, language, and thought* (2005), pp. 129–163.

[32] Roland Barthes. *Camera lucida: Reflections on photography*. London: Macmillan, 1981, pp. 54–61.

[33] Roland Barthes. "Camera Lucida: Reflections on Photography, trans. R. Howard, New York: Hill &*Wang.Orig.*" In: *La Chambre Claire, Note sur la Photographie* (1980).

[34] Andrew L Beam and Isaac S Kohane. "Big data and machine learning in health care." In: *Jama* 319.13 (2018), pp. 1317–1318.

[35] Kent Beck. *Extreme programming explained: embrace change*. addison-wesley professional, 2000.

[36] Djamila Romaissa Beddiar et al. "Vision-based human activity recognition: a survey." In: *Multimedia Tools and Applications* 79 (2020), pp. 30509–30555.

[37] Imad Eddine Ibrahim Bekkouch, Victoria Eyharabide, and Frederic Billiet. "Dual Training for Transfer Learning: Application on Medieval Studies." In: *2021 International Joint Conference on Neural Networks (IJCNN)*. Ieee. 2021, pp. 1–8.

[38] Michael van Bekkum et al. "Modular design patterns for hybrid learning and reasoning systems: a taxonomy, patterns and use cases." In: *Applied Intelligence* 51.9 (2021), pp. 6528–6546.

[39] Emma Bell and Jane Davison. "Visual management studies: Empirical and theoretical approaches." In: *International Journal of Management Reviews* 15.2 (2013), pp. 167–184.

[40] Emma Bell, Samantha Warren, and Jonathan E Schroeder. *The Routledge companion to visual organization*. Routledge London, 2014.

[41] Emily M Bender and Batya Friedman. "Data statements for natural language processing: Toward mitigating system bias and enabling better science." In: *Transactions of the Association for Computational Linguistics* 6 (2018), pp. 587–604.

[42] Philipp Benz et al. "Robustness May Be at Odds with Fairness: An Empirical Study on Class-wise Accuracy." In: *NeurIPS 2020 Workshop on Preregistration in Machine Learning, 11 December 2020, Virtual Event*. Ed. by Luca Bertinetto et al. Vol. 148. Proceedings of Machine Learning Research. Pmlr, 2020, pp. 325–342.

[43] Elena Beretta et al. "Ethical and socially-aware data labels." In: *Annual International Symposium on Information Management and Big Data*. Springer. 2018, pp. 320–327.

[44] David Berry. *David Berry: The explainability turn and Digital Humanities*. Ed. by Digital Humanities at MIT Libraries. 2021.

[45] David M. Berry. "AI, Ethics, and Digital Humanities." In: *The Bloomsbury Handbook to the Digital Humanities* (2022), p. 445.

[46] Federico Bertola and Viviana Patti. "Ontology-based affective models to organize artworks in the social semantic web." In: *Inf. Process. Manag.* 52.1 (2016), pp. 139–162. DOI: 10.1016/j.ipm.2015.10.003.

[47] Andrew Bevan. "The data deluge." In: *Antiquity* 89.348 (2015), pp. 1473–1484.

[48] Abeba Birhane. "Algorithmic Colonization of Africa." In: *SCRIPTed* 17.2 (2020).

[49] Eva Blomqvist et al. "Experimenting with eXtreme design." In: *Knowledge Engineering and Management by the Masses: 17th International Conference, EKAW 2010, Lisbon, Portugal, October 11-15, 2010. Proceedings 17.* Springer. 2010, pp. 120–134.

[50] Jorge Luis Borges. "Funes el memorioso." In: *Ficciones* (1944), pp. 519–525.

[51] Jorge Luis Borges. "The analytical language of John Wilkins." In: *Other inquisitions* 1952 (1937), pp. 101–105.

[52] Anna M. Borghi and Ferdinand Binkofski. *Words as social tools: An embodied view on abstract concepts*. Vol. 2. Springer, 2014.

[53] Anna M. Borghi et al. "Varieties of abstract concepts: development, use and representation in the brain." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1752 (2018), p. 20170121. DOI: 10.1098/rstb.2017.0121.

[54] Anna M. Borghi et al. "Words as social tools: Language, sociality and inner grounding in abstract concepts." en. In: *Physics of Life Reviews* 29 (2019), pp. 120–153. ISSN: 1571-0645. DOI: 10.1016/j.plrev.2018.12.001.

[55] Ali Borji. "Negative results in computer vision: A perspective." In: *Image and Vision Computing* 69 (2018), pp. 1–8.

[56] Lukas Bossard, Matthieu Guillaumin, and Luc Van. "Event Recognition in Photo Collections with a Stopwatch HMM." In: *2013 IEEE International Conference on Computer Vision*. Sydney, Australia: Ieee, 2013, pp. 1193–1200. DOI: 10.1109/iccv.2013.151.

[57] Kaila C. Bruer et al. "Identifying liars through automatic decoding of children's facial expressions." In: *Child development* 91.4 (2020), e995–e1011.

[58] Jerome Bruner. "Culture and human development: A new look." In: *Human development* 33.6 (1990), pp. 344–355.

[59] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. "Concreteness ratings for 40 thousand generally known English word lemmas." In: *Behavior research methods* 46 (2014), pp. 904–911.

[60] Joy Buolamwini. "Facing the Coded Gaze with Evocative Audits and Algorithmic Audits." PhD thesis. Massachusetts Institute of Technology, 2022.

[61] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In: *Conference on fairness, accountability and transparency*. Pmlr. 2018, pp. 77–91.

[62] Joy Adowaa Buolamwini. "Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers." PhD thesis. Massachusetts Institute of Technology, 2017.

[63] Judith Butler. *Bodies That Matter: On the Discursive Limits of "Sex"*. Routledge, 1993.

[64] Kirill Bykov et al. "NoiseGrad - Enhancing Explanations by Introducing Stochasticity to Model Weights." In: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 2022, pp. 6132–6140.

[65] Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. "Fatality killed the cat or: BabelPic, a multimodal dataset for non-concrete concepts." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 4680–4686. DOI: 10.18653/v1/2020.acl-main.425.

[66] Flavio Calmon et al. "Optimized pre-processing for discrimination prevention." In: *Advances in neural information processing systems* 30 (2017).

[67] Jianfang Cao, Yanfei Li, and Yun Tian. "Emotional modelling and classification of a large-scale collection of scene images in a cluster environment." en. In: *Plos One* 13.1 (2018), e0191064. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0191064.

[68] Nicola Carboni and Livio de Luca. "An ontological approach to the description of visual and iconographical representations." In: *Heritage* 2.2 (2019), pp. 1191–1210.

[69] Eva Cetinic, Tomislav Lipic, and Sonja Grgic. "A Deep Learning Perspective on Beauty, Sentiment, and Remembrance of Art." In: *IEEE Access* 7 (2019), pp. 73694–73710. ISSN: 2169-3536. DOI: 10.1109/access.2019.2921101.

[70] Eva Cetinic, Tomislav Lipic, and Sonja Grgic. "Fine-tuning Convolutional Neural Networks for fine art classification." In: *Expert Syst. Appl.* 114 (2018), pp. 107–118. DOI: 10.1016/j.eswa.2018.07.026.

[71] Eva Cetinic and James She. "Understanding and creating art with AI: Review and outlook." In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18.2 (2022), pp. 1–22.

[72] Arjun Chandrasekaran et al. "We are Humor Beings: Understanding and Predicting Visual Humor." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: Ieee, 2016, pp. 4603–4612. DOI: 10.1109/cvpr.2016.498.

[73] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. "Revolt: Collaborative crowdsourcing for labeling machine learning datasets." In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2017, pp. 2334–2346.

[74] Aditya Chattopadhyay et al. "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks." In: *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*. IEEE Computer Society, 2018, pp. 839–847. DOI: 10.1109/wacv.2018.00097.

[75] Jiawei Chen et al. "Bias and debias in recommender system: A survey and future directions." In: *ACM Transactions on Information Systems* 41.3 (2023), pp. 1–39.

[76] Leiyu Chen et al. "Review of Image Classification Algorithms Based on Convolutional Neural Networks." In: *Remote. Sens.* 13.22 (2021), p. 4712. DOI: 10.3390/rs13224712.

[77] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. "When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations." In: *International Conference on Learning Representations*. 2021.

[78] Xinlei Chen et al. "Iterative visual reasoning beyond convolutions." In: *Proc. of CVPR 2018*. Ieee. 2018, pp. 7239–7248.

[79] Xiao Chu et al. "Multi-task Recurrent Neural Network for Immediacy Prediction." en. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: Ieee, 2015, pp. 3352–3360. DOI: 10.1109/iccv.2015.383.

[80] Tat-Seng Chua et al. "NUS-WIDE: a real-world web image database from National University of Singapore." In: *Proceedings of the ACM International Conference on Image and Video Retrieval*. Civr '09. New York, NY, USA: Association for Computing Machinery, 2009, pp. 1–9. DOI: 10.1145/1646396.1646452.

[81] Ching-Yao Chuang et al. "Learning to Act Properly: Predicting and Explaining Affordances from Images." In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: Ieee, 2018, pp. 975–983. DOI: 10.1109/cvpr.2018.00108.

[82] Fiorela Ciroku et al. "Automated multimodal sensemaking: Ontology-based integration of linguistic frames and visual data." In: *Computers in Human Behavior* 150 (2024), p. 107997.

[83] Shannon Ciston. *A Critical Field Guide For Working With Machine Learning Datasets*. 2023.

[84] James N. MA Cohen and Paul Mihailidis. "Exploring Curation as a core competency in digital and media literacy education." In: *Faculty Works: Digital Humanities &NewMedia* 4 (2013).

[85] Mihai Gabriel Constantin et al. "Visual interestingness prediction: A benchmark framework and literature review." In: *International Journal of Computer Vision* 129 (2021), pp. 1526–1550.

[86] Silvia Corchs, Elisabetta Fersini, and Francesca Gasparini. "Ensemble Learning on Visual and Textual Data for Social Image Emotion Classification." In: *International Journal of Machine Learning and Cybernetics* 10.8 (2019), pp. 2057–2070. ISSN: 1868-8071, 1868-808x. DOI: 10.1007/s13042-017-0734-0.

[87] Kate Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

[88] NAC Cressie and HJ Whitford. "How to use the two sample t-test." In: *Biometrical Journal* 28.2 (1986), pp. 131–148.

[89] Marco Cristani et al. "Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis." In: *Proceedings of the 21st ACM international conference on Multimedia*. 2013, pp. 213–222.

[90] Elliot J Crowley and Andrew Zisserman. "The art of detection." In: *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I 14*. Springer. 2016, pp. 721–737.

[91] Sebastian J Crutch, Sarah Connell, and Elizabeth K Warrington. "The different representational frameworks underpinning abstract and concrete knowledge: Evidence from odd-one-out judgements." In: *Quarterly Journal of Experimental Psychology* 62.7 (2009), pp. 1377–1390. ISSN: 1747-0218. DOI: 10.1080/17470210802483834.

[92] Sebastian J Crutch, Basil H Ridha, and Elizabeth K Warrington. "The different frameworks underlying abstract and concrete knowledge: Evidence from a bilingual patient with a semantic refractory access dysphasia." In: *Neurocase* 12.3 (2006), pp. 151–163.

[93] Emely Pujólli da Silva et al. "Recognition of Affective and Grammatical Facial Expressions: A Study for Brazilian Sign Language." In: *Computer Vision – ECCV 2020 Workshops*. Ed. by Adrien Bartoli and Andrea Fusiello. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 218–236. DOI: 10.1007/978-3-030-66096-3_16.

[94] Wenliang Dai et al. "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning." In: *arXiv preprint arXiv:2305.06500* (2023). arXiv: 2305.06500 [cs.CV]. URL: https://doi.org/10.48550/arXiv.2305.06500.

[95] Roxana Daneshjou et al. "Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review." In: *JAMA dermatology* 157.11 (2021), pp. 1362–1369.

[96] Marilena Daquino and Francesca Tomasi. "Historical Context Ontology (HiCO): a conceptual model for describing context information of cultural heritage objects." In: *Research Conference on Metadata and Semantics Research*. Springer. 2015, pp. 424–436.

[97] Stamatia Dasiopoulou, Ioannis Kompatsiaris, and Michael G Strintzis. "Applying fuzzy DLs in the extraction of image semantics." In: *Journal on data semantics XIV*. Springer, 2009, pp. 105–132.

[98] Ritendra Datta et al. "Studying Aesthetics in Photographic Images Using a Computational Approach." In: *Computer Vision – ECCV 2006*. Ed. by Aleš Leonardis, Horst Bischof, and Axel Pinz. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006, pp. 288–301. DOI: 10.1007/11744078_23.

[99] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. "Racial Bias in Hate Speech and Abusive Language Detection Datasets." In: *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 25–35.

[100] Charles P. Davis and Eiling Yee. "Building semantic memory from embodied and distributional language experience." en. In: *WIREs Cognitive Science* 12.5 (2021), e1555. ISSN: 1939-5086. DOI: 10.1002/wcs.1555.

[101] Charles P. Davis and Eiling Yee. "Building semantic memory from embodied and distributional language experience." In: *WIREs Cognitive Science* e1555 (2021).

[102] S De Giorgis and A Gangemi. "Exuviae: an ontology for conceptual epistemic comparison." In: *2022 Proceedings of the 6th International Conference on Graphs and Networks in the Humanities, Amsterdam, Netherlands (2022, accepted)*. 2022.

[103] Stefano De Giorgis, Aldo Gangemi, and Rossana Damiano. "Basic Human Values and Moral Foundations Theory in ValueNet Ontology." In: *International Conference on Knowledge Engineering and Knowledge Management*. Springer. 2022, pp. 3–18.

[104] Stefano De Giorgis, Aldo Gangemi, and Dagmar Gromann. "Imageschemanet: Formalizing embodied commonsense knowledge providing an imageschematic layer to framester." In: *Semantic Web Journal, forthcoming* (2022).

[105] Jia Deng et al. "Imagenet: A large-scale hierarchical image database." In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[106] Daniel Deutch, Tanu Malik, and Adriane Chapman. "Theory and Practice of Provenance." In: *Proceedings of the 2022 International Conference on Management of Data*. 2022, pp. 2544–2545.

[107] Prafulla Dhariwal and Alexander Quinn Nichol. "Diffusion Models Beat GANs on Image Synthesis." In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. Ed. by Marc'Aurelio Ranzato et al. 2021, pp. 8780–8794.

[108] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. "Are You Really Smiling at Me? Spontaneous versus Posed Enjoyment Smiles." In: *Computer Vision – ECCV 2012*. Ed. by Andrew Fitzgibbon et al. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2012, pp. 525–538. DOI: 10.1007/978-3-642-33712-3_38.

[109] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In: *International Conference on Learning Representations*. 2020.

[110] Jon Andoni Duñabeitia et al. "Qualitative differences in the representation of abstract versus concrete words: Evidence from the visual-world paradigm." In: *Cognition* 110.2 (2009), pp. 284–292.

[111] Kevin Durda, Lori Buchanan, and Richard Caron. "Grounding co-occurrence: Identifying features in a lexical co-occurrence model of semantic memory." In: *Behavior Research Methods* 41.4 (2009), pp. 1210–1223.

[112] John P Eakins. "Retrieval of still images by content." In: *European Summer School on Information Retrieval*. Springer. 2000, pp. 111–138.

[113] Jim Edwards. *We are Now Posting a Staggering 1.8 Billion Photos to Social Media Every Day*. 2014.

[114] Paul Ekman et al. "Basic emotions." In: *Handbook of cognition and emotion* 98.45-60 (1999), p. 16.

[115] Peter Enser and Peter Enser. *Visual image retrieval: seeking the alliance of concept-based and content-based paradigms*. 1999.

[116] D Erhan et al. "Visualizing higher-layer features of a deep network." In: *University of Montreal* 1341 (2009).

[117] Dumitru Erhan et al. "Visualizing higher-layer features of a deep network." In: *University of Montreal* 1341.3 (2009), p. 1.

[118] Sergio Escalera et al. "ChaLearn Looking at People 2015: Apparent Age and Cultural Event Recognition Datasets and Results." In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. Santiago, Chile: Ieee, 2015, pp. 243–251. DOI: 10.1109/iccvw.2015.40.

[119] Charles J Fillmore. "Frame semantics." In: *Linguistics in the Morning Calm*. Seoul: Hanshin, 1982, pp. 111–138.

[120] Charles J Fillmore. "Frames and the semantics of understanding." In: *Quaderni di semantica* 6.2 (1985), pp. 222–254.

[121] Chaz Firestone and Brian J Scholl. "Cognition does not affect perception: Evaluating the evidence for "top-down" effects." In: *Behavioral and brain sciences* 39 (2016).

[122] J. R. Firth. "A synopsis of linguistic theory, 1930-1955." In: *Studies in Linguistic Analysis* (1957).

[123] Susan T Fiske and Shelley E Taylor. *Social cognition*. Mcgraw-Hill Book Company, 1991.

[124] Hal Foster. *Preface. Vision and Visuality. Ed. Hal Foster*. 1988.

[125] Sorelle A Friedler et al. "A comparative study of fairness-enhancing interventions in machine learning." In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 329–338.

[126] Ruigang Fu et al. "Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs." In: *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

[127] Andrew C. Gallagher and Tsuhan Chen. "Understanding Images of Groups of People." In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL: Ieee, 2009, pp. 256–263. DOI: 10.1109/cvpr.2009.5206828.

[128] Aldo Gangemi. "Closing the Loop between knowledge patterns in cognition and the Semantic Web." In: *Semantic Web* 11.1 (2020), pp. 139–151.

[129] Aldo Gangemi and Peter Mika. "Understanding the semantic web through descriptions and situations." In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2003, pp. 689–706.

[130] Aldo Gangemi, Valentina Presutti, and Mehwish Alam. "Amnestic Forgery: An Ontology of Conceptual Metaphors." In: *Formal Ontology in Information Systems - Proceedings of the 10th International Conference, FOIS 2018, Cape Town, South Africa, 19-21 September 2018*. IOS Press, 2018.

[131] Aldo Gangemi et al. "Framester: A wide coverage linguistic linked data hub." en. In: *European Knowledge Acquisition Workshop*. Ed. by Eva Blomqvist et al. Lecture Notes in Computer Science. Springer. Cham: Springer International Publishing, 2016, pp. 239–254. DOI: 10.1007/978-3-319-49004-5_16.

[132] Aldo Gangemi et al. "Semantic web machine reading with FRED." In: *Semantic Web* 8.6 (2017), pp. 873–893.

[133] Aldo Gangemi et al. "Sweetening ontologies with DOLCE." In: *International Conference on Knowledge Engineering and Knowledge Management*. Vol. 2473. Springer. 2002, pp. 166–181. DOI: 10.1007/3-540-45810-7_18.

[134] Aldo Gangemi et al. "Sweetening wordnet with dolce." In: *AI magazine* 24.3 (2003), pp. 13–13.

[135] Noa Garcia and George Vogiatzis. "How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval." In: 2018, pp. 0–0.

[136] Timnit Gebru et al. "Datasheets for datasets." In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[137] R Stuart Geiger et al. "Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?" In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 325–336.

[138] R Stuart Geiger et al. ""Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data?" In: *Quantitative Science Studies* 2.3 (2021), pp. 795–827.

[139] Spandana Gella, Desmond Elliott, and Frank Keller. "Cross-lingual Visual Verb Sense Disambiguation." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 1998–2004.

[140] Spandana Gella, Mirella Lapata, and Frank Keller. "Unsupervised Visual Sense Disambiguation for Verbs using Multimodal Embeddings." In: *Proceedings of NAACL-HLT*. 2016, pp. 182–192.

[141] Dedre Gentner and Jennifer Asmuth. "Metaphoric extension, relational categories, and abstraction." en. In: *Language, Cognition and Neuroscience* 34.10 (2019), pp. 1298–1307. ISSN: 2327-3798, 2327-3801. DOI: 10.1080/23273798.2017.1410560.

[142] Bhavya Ghai and Klaus Mueller. "D-BIAS: a causality-based human-in-the-loop system for tackling algorithmic bias." In: *IEEE Transactions on Visualization and Computer Graphics* 29.1 (2022), pp. 473–482.

[143] Shreya Ghosh and Abhinav Dhall. "Role of Group Level Affect to Find the Most Influential Person in Images." In: *Computer Vision – ECCV 2018 Workshops*. Ed. by Laura Leal-Taixé and Stefan Roth. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 518–533. DOI: 10.1007/978-3-030-11012-3_39.

[144] Jacob Gildenblat and contributors. *PyTorch library for CAM methods*. https://github.com/jacobgil/pytorch-grad-cam. 2021.

[145] August Corrons Giménez and Lluís Garay Tamajón. "Analysis of the third-order structuring of Shalom Schwartz's theory of basic human values." In: *Heliyon* 5.6 (2019), e01797.

[146] Gina Giotta. "Ways of seeing... what you want: flexible visuality and image politics in the post-truth era." In: *Fake News: Understanding Media and Misinformation in the Digital Age* (2020), p. 29.

[147] Arushi Goel, Keng Teck Ma, and Cheston Tan. "An End-To-End Network for Generating Social Relationship Graphs." In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: Ieee, 2019, pp. 11178–11187. DOI: 10.1109/cvpr.2019.01144.

[148] Jesse Graham et al. "Moral foundations theory: The pragmatic validity of moral pluralism." In: *Advances in experimental social psychology*. Vol. 47. Elsevier, 2013, pp. 55–130.

[149] Douglas Gray et al. "Predicting Facial Beauty without Landmarks." In: *Computer Vision – ECCV 2010*. Ed. by Kostas Daniilidis, Petros Maragos, and Nikos Paragios. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2010, pp. 434–447. DOI: 10.1007/978-3-642-15567-3_32.

[150] Howard Greisdorf and Brian O'Connor. "Modelling what users see when they look at images: a cognitive viewpoint." In: *Journal of Documentation* 58.1 (2002), pp. 6–29. ISSN: 0022-0418. DOI: 10.1108/00220410210425386.

[151] Tom Gruber. *What is an Ontology*. 1993.

[152] Jiuxiang Gu et al. "Scene Graph Generation With External Knowledge and Image Reconstruction." In: 2019, pp. 1969–1978.

[153] Nicola Guarino, Daniel Oberle, and Steffen Staab. "What is an ontology?" In: *Handbook on ontologies* (2009), pp. 1–17.

[154] Meiqi Guo, Rebecca Hwa, and Adriana Kovashka. "Detecting Persuasive Atypicality by Modeling Contextual Compatibility." In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: Ieee, 2021, pp. 952–962. DOI: 10.1109/iccv48922.2021.00101.

[155] Shu Guo et al. "Jointly Embedding Knowledge Graphs and Logical Rules." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, 2016, pp. 192–202. DOI: 10.18653/v1/D16-1019.

[156] Wenzhong Guo, Jianwen Wang, and Shiping Wang. "Deep multimodal representation learning: A survey." In: *IEEE Access* 7 (2019), pp. 63373–63394.

[157] Robert M Haralick and Linda G Shapiro. "Glossary of computer vision terms." In: *Pattern Recognit.* 24.1 (1991), pp. 69–93.

[158] Donna Haraway. "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective." In: *Feminist Studies* 14.3 (1988), pp. 575–599. DOI: 10.2307/3178066.

[159] Jonathon S Hare et al. "Mind the gap: another look at the problem of the semantic gap in image retrieval." In: *Multimedia Content Analysis, Management, and Retrieval 2006*. Ed. by Edward Y Chang, Alan Hanjalic, and Nicu Sebe. Vol. 6073. International Society for Optics and Photonics. San Jose, CA: Spie, 2006, p. 607309. DOI: 10.1117/12.647755.

[160] Marcel Harpaintner, Natalie M. Trumpp, and Markus Kiefer. "The Semantic Content of Abstract Concepts: A Property Listing Study of 296 Abstract Words." In: *Frontiers in Psychology* 9 (2018), p. 1748. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.01748.

[161] Catherine Havasi, Robert Speer, and Jason Alonso. "ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge." In: *Recent advances in natural language processing*. John Benjamins Philadelphia, PA. 2007, pp. 27–29.

[162] Gregor U Hayn-Leichsenring, Thomas Lehmann, and Christoph Redies. "Subjective ratings of beauty and aesthetics: correlations with statistical image properties in western oil paintings." In: *i-Perception* 8.3 (2017), p. 2041669517715474.

[163] Kaiming He et al. "Deep residual learning for image recognition." en. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Las Vegas, NV, USA: Ieee, 2016, pp. 770–778. DOI: 10.1109/cvpr.2016.90.

[164] Kaiming He et al. "Mask r-cnn." In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[165] Pascal Hitzler and Md Kamruzzaman Sarker. *Neuro-symbolic artificial intelligence: The state of the art*. 2022.

[166] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models." In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle et al. Vol. 33. 2020, pp. 6840–6851.

[167] Paul Hoffman. "Concepts, control, and context: A connectionist account of normal and disordered semantic cognition." en. In: *Psychological Review* 125.3 (2018), p. 293. ISSN: 1939-1471. DOI: 10.1037/rev0000094.

[168] Derek Hoiem, Alexei A Efros, and Martial Hebert. "Putting objects in perspective." In: *International Journal of Computer Vision* 80 (2008), pp. 3–15.

[169] Sarah Holland et al. "The dataset nutrition label." In: *Data Protection and Privacy* 12.12 (2020), p. 1.

[170] Markus A Höllerer, Dennis Jancsary, and Maria Grafström. "'A picture is worth a thousand words': Multimodal sensemaking of the global financial crisis." In: *Organization Studies* 39.5-6 (2018), pp. 617–644.

[171] Wei-Lin Hsiao and Kristen Grauman. "Learning the Latent "Look": Unsupervised Discovery of a Style-Coherent Embedding from Fashion Images." In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: Ieee, 2017, pp. 4213–4222. DOI: 10.1109/iccv.2017.451.

[172] X. Huang and A. Kovashka. "Inferring Visual Persuasion via Body Language, Setting, and Deep Features." In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 2016, pp. 778–784. DOI: 10.1109/cvprw.2016.102.

[173] Hayley Hung and Daniel Gatica-Perez. "Estimating cohesion in small groups using audio-visual nonverbal behavior." In: *IEEE Transactions on Multimedia* 12.6 (2010), pp. 563–575.

[174] Zaeem Hussain et al. "Automatic Understanding of Image and Video Advertisements." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1705–1715.

[175] Rami Ibrahim and M. Omair Shafiq. "Explainable Convolutional Neural Networks: A Taxonomy, Review, and Future Directions." In: *ACM Comput. Surv.* 55.10 (2023). ISSN: 0360-0300. DOI: 10.1145/3563691.

[176] Filip Ilievski, Pedro Szekely, and Bin Zhang. "Cskg: The commonsense knowledge graph." en. In: *European Semantic Web Conference*. Ed. by Ruben Verborgh et al. Lecture Notes in Computer Science. Springer. Cham: Springer International Publishing, 2021, pp. 680–696. DOI: 10.1007/978-3-030-77385-4_41.

[177] Leonardo Impett and Franco Moretti. *Totentanz. operationalizing aby warburg's pathosformeln*. Tech. rep. Stanford Literary Lab, 2017.

[178] *Instagram - Claire Barrow*. https://www.instagram.com/claire_barrow/.

[179] *Instagram - Mikey Joyce*. https://www.instagram.com/m___joyce/.

[180] Phillip Isola, Joseph J Lim, and Edward H Adelson. "Discovering states and transformations in image collections." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1383–1391.

[181] Johan Jansson and Brian J Hracs. "Conceptualizing curation in the age of abundance: The case of recorded music." In: *Environment and Planning A: Economy and Space* 50.8 (2018), pp. 1602–1625.

[182] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. "Diffusion Models for Counterfactual Explanations." In: *CoRR* abs/2203.15636 (2022). DOI: 10.48550/arXiv.2203.15636.

[183] Menglin Jia et al. "Intentonomy: a Dataset and Study towards Human Intent Understanding." In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: Ieee, 2021, pp. 12981–12991. DOI: 10.1109/cvpr46437.2021.01279.

[184] Eun Seo Jo and Timnit Gebru. "Lessons from archives: Strategies for collecting sociocultural data in machine learning." In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 306–316.

[185] Brendan T Johns and Michael N Jones. "Perceptual inference through global lexical similarity." In: *Topics in Cognitive Science* 4.1 (2012), pp. 103–120.

[186] Jungseock Joo, Francis F. Steen, and Song-Chun Zhu. "Automated Facial Trait Judgment and Election Outcome Prediction: Social Dimensions of Face." In: *2015 IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: Ieee, 2015, pp. 3712–3720. DOI: 10.1109/iccv.2015.423.

[187] Jungseock Joo et al. "Visual Persuasion: Inferring Communicative Intents of Images." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 216–223.

[188] Corinne Jörgensen. *Image Retrieval: Theory and Research*. en. Scarecrow Press, 2003.

[189] Hyungsik Jung and Youngrock Oh. "Towards Better Explanations of Class Activation Mapping." In: *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. Ieee, 2021, pp. 1316–1324. DOI: 10.1109/iccv48922.2021.00137.

[190] David Kadish, Sebastian Risi, and Anders Sundnes Løvlie. "Improving object detection in art images using only style transfer." In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, pp. 1–8.

[191] Nasrin Kalanat and Adriana Kovashka. "Symbolic image detection using scene and knowledge graphs." In: *arXiv preprint arXiv:2206.04863* (2022).

[192] Ibrahem Kandel and Mauro Castelli. "How deeply to fine-tune a convolutional neural network: a case study using a histopathology dataset." In: *Applied Sciences* 10.10 (2020), p. 3359.

[193] Imannuel Kant. *Critique of pure reason. 1781. Modern Classical Philosophers*. 1908.

[194] Alex Kendall and Roberto Cipolla. "Geometric loss functions for camera pose regression with deep learning." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5974–5983.

[195] Daniel Keren. "Painter identification using local features and naive bayes." In: *2002 International Conference on Pattern Recognition*. Vol. 2. IEEE. 2002, pp. 474–477.

[196] Abdullah Khalili and Hamid Bouchachia. "An information theory approach to aesthetic assessment of visual patterns." In: *Entropy* 23.2 (2021), p. 153.

[197] Pooja Kherwa and Poonam Bansal. "Topic Modeling: A Comprehensive Review." In: *EAI Endorsed Trans. Scalable Inf. Syst.* 7.24 (2020), e2. DOI: 10.4108/eai.13-7-2018.159623.

[198] Aditya Khosla et al. "Looking Beyond the Visible Scene." In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: Ieee, 2014, pp. 3710–3717. DOI: 10.1109/cvpr.2014.474.

[199] M. Hadi Kiapour et al. "Hipster Wars: Discovering Elements of Fashion Styles." In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 472–488. DOI: 10.1007/978-3-319-10590-1_31.

[200] Douwe Kiela and Léon Bottou. "Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 36–45. DOI: 10.3115/v1/D14-1005.

[201] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization." In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015.

[202] Marijn Koolen, Jasmijn Van Gorp, and Jacco Van Ossenbruggen. "Toward a model for digital tool criticism: Reflection as integrative practice." In: *Digital Scholarship in the Humanities* 34.2 (2019), pp. 368–385.

[203] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. "Do Better ImageNet Models Transfer Better?" In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 2661–2671. DOI: 10.1109/cvpr.2019.00277.

[204] Jean Kossaifi et al. "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild." In: *IEEE transactions on pattern analysis and machine intelligence* 43.3 (2019), pp. 1022–1040.

[205] Stavroula-Thaleia Kousta et al. "The representation of abstract words: why emotion matters." In: *Journal of Experimental Psychology: General* 140.1 (2011), p. 14. ISSN: 1939-2222. DOI: 10.1037/a0021446.

[206] D Krech. "Rdflib: A python library for working with rdf." In: *Online https://github.com/RDFLib/rdflib* (2006).

[207] Gabriel Kreiman. *Neuroscience: Literary inspiration*. 2011.

[208] Ranjay Krishna et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." In: *International journal of computer vision* 123.1 (2017), pp. 32–73.

[209] Ranjay Krishna et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." In: *International journal of computer vision* 123.1 (2017), pp. 32–73.

[210] Alex Krizhevsky and Geoff Hinton. "Convolutional deep belief networks on cifar-10." In: *Unpublished manuscript* 40.7 (2010), pp. 1–9.

[211] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." In: *Advances in neural information processing systems* 25 (2012).

[212] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[213] Alina Kuznetsova et al. "The Open Images Dataset V4." en. In: *International Journal of Computer Vision* 128.7 (2020), pp. 1956–1981. ISSN: 1573-1405. DOI: 10.1007/s11263-020-01316-z.

[214] Camila Laranjeira, Virgínia Fernandes Mota, and Jefferson Alex dos Santos. "Machine Learning Bias in Computer Vision: Why do I have to care?" 2021.

[215] Angeliki Lazaridou et al. "Combining Language and Vision with a Multimodal Skip-gram Model." In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. 2015.

[216] Dieu-Thu Le, Jasper Uijlings, and Raffaella Bernardi. "Exploiting language models for visual recognition." In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013, pp. 769–779.

[217] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." In: *Nature* 521.7553 (2015), pp. 436–444.

[218] Isabella di Lenardo, Benoît Laurent Auguste Seguin, and Frédéric Kaplan. *Visual patterns discovery in large databases of paintings*. Tech. rep. 2016.

[219] Guang Li et al. "Entangled Transformer for Image Captioning." In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): Ieee, 2019, pp. 8927–8936. DOI: 10.1109/iccv.2019.00902.

[220] Junnan Li et al. "Dual-Glance Model for Deciphering Social Relationships." In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: Ieee, 2017, pp. 2669–2678. DOI: 10.1109/iccv.2017.289.

[221] Junnan Li et al. "Visual Social Relationship Recognition." en. In: *International Journal of Computer Vision* 128.6 (2020), pp. 1750–1764. ISSN: 1573-1405. DOI: 10.1007/s11263-020-01295-1.

[222] Ruiyu Li et al. "Situation Recognition with Graph Neural Networks." In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: Ieee, 2017, pp. 4183–4192. DOI: 10.1109/iccv.2017.448.

[223] Wanhua Li et al. "Graph-Based Social Relation Reasoning." In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 18–34. DOI: 10.1007/978-3-030-58555-6_2.

[224] Yujia Li et al. "Gated graph sequence neural networks." In: *International Conference on Learning Representations*. 2016.

[225] Peiyuan Liao et al. "The artbench dataset: Benchmarking generative models with artworks." In: *arXiv preprint arXiv:2206.11404* (2022).

[226] Baihan Lin. "Knowledge Management System with NLP-Assisted Annotations: A Brief Survey and Outlook." In: *Proceedings of the CIKM 2022 Workshops co-located with 31st ACM International Conference on Information and Knowledge Management (CIKM 2022), Atlanta, USA, October 17-21, 2022*. Ed. by Georgios Drakopoulos and Eleanna Kafeza. Vol. 3318. CEUR Workshop Proceedings. CEUR-WS.org, 2022.

[227] Tianyang Lin et al. "A survey of transformers." In: *AI Open* 3 (2022), pp. 111–132. DOI: 10.1016/j.aiopen.2022.10.001.

[228] Hugo Liu and Push Singh. "ConceptNet–a practical commonsense reasoning tool-kit." In: *BT technology journal* 22.4 (2004), pp. 211–226.

[229] Mengyi Liu et al. "Exploiting Feature Hierarchies with Convolutional Neural Networks for Cultural Event Recognition." In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. Santiago, Chile: Ieee, 2015, pp. 274–279. DOI: 10.1109/iccvw.2015.44.

[230] Jose Llamas et al. "Applying deep learning techniques to cultural heritage images within the inception project." In: *Euro-Mediterranean Conference*. Springer. 2016, pp. 25–32.

[231] Ben London et al. "Collective activity detection using hinge-loss Markov random fields." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2013, pp. 566–571.

[232] Max M. Louwerse. "Knowing the Meaning of a Word by the Linguistic and Perceptual Company It Keeps." en. In: *Topics in Cognitive Science* 10.3 (2018), pp. 573–589. ISSN: 1756-8765. DOI: 10.1111/tops.12349.

[233] Cewu Lu et al. "Visual relationship detection with language priors." In: *Proc. of ECCV 2016*. Springer. 2016, pp. 852–869.

[234] Fan Lyu et al. "Attend and imagine: Multi-label image classification with visual attention and recurrent neural networks." In: *IEEE Transactions on Multimedia* 21.8 (2019), pp. 1971–1981.

[235] Elena Madison. "Frans de Waal, Good Natured: The Origins of Right and Wrong in Humans and Other Animals. Cambridge and London: Harvard University Press, 1996." In: *Cultural Logic: A Journal of Marxist Theory &Practice* 9 (2002).

[236] Aravindh Mahendran and Andrea Vedaldi. "Visualizing Deep Convolutional Neural Networks Using Natural Pre-images." In: *Int. J. Comput. Vis.* 120.3 (2016), pp. 233–255. DOI: 10.1007/s11263-016-0911-8.

[237] Gary Marcus. "The next decade in AI: four steps towards robust artificial intelligence." In: *arXiv preprint arXiv:2002.06177* (2020).

[238] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. "The More You Know: Using Knowledge Graphs for Image Classification." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 20–28. DOI: 10.1109/cvpr.2017.10.

[239] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. "The More You Know: Using Knowledge Graphs for Image Classification." In: (2017), pp. 20–28. DOI: 10.1109/cvpr.2017.10.

[240] D. S. Martinez Pandiani and V. Presutti. "Automatic Modeling of Social Concepts Evoked by Art Images as Multimodal Frames." In: *Proceedings of the Workshops and Tutorials held at LDK 2021 co-located with the 3rd Language, Data and Knowledge Conference (LDK 2021)* (2021).

[241] D.S. Martinez Pandiani and V. Presutti. "Coded Visions: Addressing Cultural Bias in Image Annotation Systems with the Descriptions and Situations Ontology Design Pattern." In: *6th International Conference of Graphs and Networks in the Humanities 2022: Technologies, Models, Analyses, and Visualizations* (2022).

[242] D.S. Martinez Pandiani and V. Presutti. "Situated Ground Truths: Enhancing Bias-Aware AI by Situating Data Labels with SituAnnotate." In: *Special Issue on Trustworthy Artificial Intelligence of ACM Transactions on Knowledge Discovery from Data (TKDD)* (2024).

[243] D.S. Martinez Pandiani et al. "Hypericons for interpretability: decoding abstract concepts in visual data." In: *International Journal of Digital Humanities* (2023), pp. 1–40.

[244] David Masip Rodo, Alexander Todorov, and Jordi Vitrià Marca. "The Role of Facial Regions in Evaluating Social Dimensions." en. In: *Computer Vision – ECCV 2012. Workshops and Demonstrations*. Ed. by Andrea Fusiello, Vittorio Murino, and Rita Cucchiara. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2012, pp. 210–219. DOI: 10.1007/978-3-642-33868-7_21.

[245] Claudio Masolo et al. "Social roles and their descriptions." In: *Proceedings of the Ninth International Conference on Principles of Knowledge Representation and Reasoning*. 2004, pp. 267–277.

[246] José Maurício, Inês Domingues, and Jorge Bernardino. "Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review." In: *Applied Sciences* 13.9 (2023), p. 5521.

[247] Claudia Mazzuca. "Queering Abstract Concepts: A Grounded Perspective on Gender." PhD thesis. Università di Bologna, 2020.

[248] Ninareh Mehrabi et al. "A survey on bias and fairness in machine learning." In: *ACM computing surveys (CSUR)* 54.6 (2021), pp. 1–35.

[249] Marco Miceli, Juan Posada, and Tantam Yang. "Studying up machine learning data: Why talk about bias when we mean power?" In: *Proceedings of the ACM on Human-Computer Interaction* 6 (2022), pp. 1–14.

[250] George A Miller. "WordNet: a lexical database for English." In: *Communications of the ACM* 38.11 (1995), pp. 39–41.

[251] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[252] Margaret Mitchell et al. "Model cards for model reporting." In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 220–229.

[253] WJ Thomas Mitchell. *Picture theory: Essays on verbal and visual representation*. University of Chicago Press, 1995.

[254] Brent Daniel Mittelstadt et al. "The ethics of algorithms: Mapping the debate." en. In: *Big Data Soc.* 3.2 (2016), p. 205395171667967.

[255] Yujian Mo et al. "Review the state-of-the-art technologies of semantic segmentation based on deep learning." In: *Neurocomputing* 493 (2022), pp. 626–646. DOI: 10.1016/j.neucom.2022.01.005.

[256] Sana Mohamed, Maurice T Png, and William Isaac. "Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence." In: *Philosophy &Technology* 33 (2020), pp. 659–684.

[257] Saif Mohammad and Svetlana Kiritchenko. "WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018.

[258] Saif M. Mohammad and Svetlana Kiritchenko. "WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. Ed. by Nicoletta Calzolari et al. European Language Resources Association (ELRA), 2018.

[259] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. "Affectnet: A database for facial expression, valence, and arousal computing in the wild." In: *IEEE Transactions on Affective Computing* 10.1 (2017), pp. 18–31.

[260] Luc Moreau et al. "The open provenance model: An overview." In: *Provenance and Annotation of Data and Processes: Second International Provenance and Annotation Workshop, IPAW 2008, Salt Lake City, UT, USA, June 17-18, 2008. Revised Selected Papers 2*. Springer. 2008, pp. 323–326.

[261] Luca Moschella et al. "Relative representations enable zero-shot latent space communication." In: *The Eleventh International Conference on Learning Representations*. 2022.

[262] Roberto Navigli and Simone Paolo Ponzetto. "BabelNet: Building a very large multilingual semantic network." In: *Proceedings of the 48th annual meeting of the association for computational linguistics*. 2010, pp. 216–225.

[263] Roberto Navigli and Simone Paolo Ponzetto. "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network." en. In: *Artificial intelligence* 193 (2012), pp. 217–250. ISSN: 0004-3702. DOI: 10.1016/j.artint.2012.07.001.

[264] Hongwei Ng et al. "Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning." In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015*. Ed. by Zhengyou Zhang et al. Acm, 2015, pp. 443–449. DOI: 10.1145/2818346.2830593.

[265] Anh Nguyen, Jason Yosinski, and Jeff Clune. "Understanding Neural Networks via Feature Visualization: A Survey." In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by Wojciech Samek et al. Vol. 11700. Lecture Notes in Computer Science. Springer, 2019, pp. 55–76. DOI: 10.1007/978-3-030-28954-6_4.

[266] Anh Mai Nguyen et al. "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks." In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. Ed. by Daniel D. Lee et al. 2016, pp. 3387–3395.

[267] Nanne van Noord. "A survey of computational methods for iconic image analysis." In: *Digital Scholarship in the Humanities* 37.4 (2022), pp. 1316–1338.

[268] Nanne van Noord and Eric O. Postma. "Learning scale-variant and scale-invariant features for deep image classification." In: *Pattern Recognit.* 61 (2017), pp. 583–592. DOI: 10.1016/j.patcog.2016.06.005.

[269] Eirini Ntoutsi et al. "Bias in data-driven artificial intelligence systems–An introductory survey." In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.3 (2020), e1356.

[270] Andrea Giovanni Nuzzolese, Aldo Gangemi, and Valentina Presutti. "Gathering lexical linked data and knowledge patterns from FrameNet." In: *Proceedings of the sixth international conference on Knowledge capture*. Acm. 2011, pp. 41–48.

[271] Fabian Offert. "Images of Image Machines. Visual Interpretability in Computer Vision for Art." In: *Computer Vision–ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part II 15*. Springer. 2018, pp. 0–0.

[272] Fabian Offert and Peter Bell. "Perceptual bias and technical metapictures: critical machine vision as a humanities challenge." In: *Ai &Society* 36 (2021), pp. 1133–1144.

[273] Fabian Offert and Peter Bell. "Understanding Perceptual Bias in Machine Vision Systems." In: *Informatik 2020* (2021).

[274] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. "Feature Visualization." In: *Distill* (2017). https://distill.pub/2017/feature-visualization. DOI: 10.23915/distill.00007.

[275] Niall O'Mahony et al. "Deep learning vs. traditional computer vision." In: *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*. Springer. 2020, pp. 128–144.

[276] Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato. "Survey on Visual Sentiment Analysis." In: *IET Image Processing* 14.8 (2020), pp. 1440–1456. ISSN: 1751-9667, 1751-9667. DOI: 10.1049/iet-ipr.2019.1270.

[277] Andrew Ortony, Gerald L Clore, and Allan Collins. *The cognitive structure of emotions*. Cambridge university press, 2022.

[278] Erwin Panofsky and Benjamin Drechsel. *Meaning in the visual arts*. University of Chicago Press Chicago, 1955.

[279] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[280] Sarah Pratt et al. "Grounded situation recognition." In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Ed. by Andrea Vedaldi et al. Lecture Notes in Computer Science. Springer. Cham: Springer International Publishing, 2020, pp. 314–332. DOI: 10.1007/978-3-030-58548-8_19.

[281] Valentina Presutti, Francesco Draicchio, and Aldo Gangemi. "Knowledge Extraction Based on Discourse Representation Theory and Linguistic Frames." In: *Knowledge Engineering and Knowledge Management - 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings*. Ed. by Annette ten Teije et al. Vol. 7603. Lecture Notes in Computer Science. Springer, 2012, pp. 114–129. DOI: 10.1007/978-3-642-33876-2_12.

[282] Valentina Presutti et al. "eXtreme design with content ontology design patterns." In: *Proc. Workshop on Ontology Patterns*. 2009, pp. 83–97.

[283] Junfei Qiu et al. "A survey of machine learning for big data processing." In: *EURASIP Journal on Advances in Signal Processing* 2016 (2016), pp. 1–16.

[284] Rodrigo Quian Quiroga. *Borges and memory: Encounters with the human brain*. Mit Press, 2012.

[285] Rodrigo Quian Quiroga. *The Forgetting Machine: Memory, Perception, and the Jennifer Aniston Neuron*. United States: BenBella Books, 2017.

[286] Pauline Rafferty and Rob Hidderley. *Indexing Multimedia and Creative Works: The Problems of Meaning and Interpretation*. London: Routledge, 2016. DOI: 10.4324/9781315252469.

[287] Maithra Raghu et al. "Do Vision Transformers See Like Convolutional Neural Networks?" In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. Ed. by Marc'Aurelio Ranzato et al. 2021, pp. 12116–12128.

[288] Rahul Raguram and Svetlana Lazebnik. "Computing iconic summaries of general visual concepts." In: *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Anchorage, AK, USA: Ieee, 2008, pp. 1–8. DOI: 10.1109/cvprw.2008.4562959.

[289] Aditya Ramesh et al. "Hierarchical Text-Conditional Image Generation with CLIP Latents." In: *CoRR* abs/2204.06125 (2022). DOI: 10.48550/arXiv.2204.06125. arXiv: 2204.06125.

[290] Muhammad Ramzan et al. "A review on state-of-the-art violence detection techniques." In: *IEEE Access* 7 (2019), pp. 107560–107575.

[291] Joseph Redmon and Ali Farhadi. "YOLO9000: better, faster, stronger." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7263–7271.

[292] Joseph Redmon et al. "You only look once: Unified, real-time object detection." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[293] Gérard Régimbeau. "Image source criticism in the age of the digital humanities." In: vol. 4. LIT Verlag Münster, 2014.

[294] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." In: *Advances in neural information processing systems* 28 (2015).

[295] Matthew Richardson and Pedro Domingos. "Markov logic networks." In: *Machine learning* 62.1-2 (2006), pp. 107–136.

[296] Daniel Riffe et al. *Analyzing media messages: Using quantitative content analysis in research*. Routledge, 2019.

[297] Horst Rittel. "Wicked problems." In: *Management Science,(December 1967)* 4.14 (1967).

[298] Nuria Rodríguez-Ortega. "Image processing and computer vision in the field of art history." In: *The Routledge Companion to Digital Humanities and Art History*. New York : Routledge, 2020: Routledge, 2020, pp. 338–357.

[299] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[300] Steven C Rosenbaum. *Curation nation*. McGraw-Hill, 2011.

[301] Candace Ross, Boris Katz, and Andrei Barbu. "Measuring Social Biases in Grounded Vision and Language Embeddings." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021, pp. 998–1008.

[302] Rasmus Rothe, Radu Timofte, and Luc Van Gool. "DLDR: Deep Linear Discriminative Retrieval for Cultural Event Classification from a Single Image." In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. Santiago, Chile: Ieee, 2015, pp. 295–302. DOI: 10.1109/iccvw.2015.47.

[303] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge." In: *Int. J. Comput. Vis.* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[304] Martinez Pandiani D. S. and Presutti V. "Automatic Modeling of Social Concepts Evoked by Art Images as Multimodal Frames." In: *Proceedings of the Workshops and Tutorials held at LDK 2021 co-located with the 3rd Language, Data and Knowledge Conference (LDK 2021)*. Zaragoza, Spain, 2021.

[305] Matthia Sabatelli et al. "Advances in Digital Music Iconography: Benchmarking the detection of musical instruments in unrestricted, non-photorealistic images from the artistic domain." In: *Digital Humanities Quarterly* 15.1 (2021).

[306] Mohammad Amin Sadeghi and Ali Farhadi. "Recognition using visual phrases." In: *Cvpr 2011*. Ieee. 2011, pp. 1745–1752.

[307] Nabile M Safdar, John D Banja, and Carolyn C Meltzer. "Ethical considerations in artificial intelligence." In: *European journal of radiology* 122 (2020), p. 108768.

[308] Lou Safra et al. "Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings." en. In: *Nature Communications* 11.1 (2020), p. 4728. ISSN: 2041-1723. DOI: 10.1038/s41467-020-18566-7.

[309] Christoph Sager, Christian Janiesch, and Patrick Zschech. "A survey of image labelling for computer vision applications." In: *Journal of Business Analytics* (2021), pp. 1–20.

[310] Sumit Saha. *A Comprehensive Guide to Convolutional Neural Networks, the Eli5 Way*. Medium, Towards Data Science. Dec. 15, 2018. URL: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

[311] Emad Isa Saleh. "Image embedded metadata in cultural heritage digital collections on the web: An analytical study." In: *Library Hi Tech* (2018).

[312] Babak Salimi, Bill Howe, and Dan Suciu. "Database repair meets algorithmic fairness." In: *ACM SIGMOD Record* 49.1 (2020), pp. 34–41.

[313] Surender Reddy Salkuti. "A survey of big data and machine learning." In: *International Journal of Electrical &ComputerEngineering*(2088 − 8708) 10.1 (2020).

[314] Haitham Samih et al. "Semantic Graph Representation and Evaluation for Generated Image Annotations." In: *International Conference on Advanced Machine Learning Technologies and Applications*. Springer. 2021, pp. 369–384.

[315] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. "How computers see gender: An evaluation of gender classification in commercial facial analysis services." In: *Proceedings of the ACM on Human-Computer Interaction* 3.Cscw (2019), pp. 1–33.

[316] Karin Kipper Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania, 2005.

[317] Candice Schumann et al. "A step toward more inclusive people annotations for fairness." In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 916–925.

[318] Shalom H Schwartz. "An overview of the Schwartz theory of basic values." In: *Online readings in Psychology and Culture* 2.1 (2012), pp. 2307–0919.

[319] Shalom H Schwartz et al. "Extending the cross-cultural validity of the theory of basic human values with a different method of measurement." In: *Journal of cross-cultural psychology* 32.5 (2001), pp. 519–542.

[320] William A Scott. "Empirical assessment of values and ideologies." In: *American Sociological Review* (1959), pp. 299–310.

[321] Cristina Segalin, Dong Seon Cheng, and Marco Cristani. "Social Profiling through Image Understanding: Personality Inference Using Convolutional Neural Networks." In: *Computer Vision and Image Understanding*. Image and Video Understanding in Big Data 156 (2017), pp. 34–50. ISSN: 1077-3142. DOI: 10.1016/j.cviu.2016.10.013.

[322] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." In: *Int. J. Comput. Vis.* 128.2 (2020), pp. 336–359. DOI: 10.1007/s11263-019-01228-7.

[323] Lior Shamir et al. "Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art." In: *ACM Transactions on Applied Perception (TAP)* 7.2 (2010), pp. 1–17.

[324] Ming Shao, Liangyue Li, and Yun Fu. "What Do You Do? Occupation Recognition in a Photo via Social Context." In: *2013 IEEE International Conference on Computer Vision*. Sydney, Australia: Ieee, 2013, pp. 3631–3638. DOI: 10.1109/iccv.2013.451.

[325] Henry L Shapiro. "Memory and Meaning: Borges and" Funes el memorioso"." In: *Revista Canadiense de Estudios Hispánicos* (1985), pp. 257–265.

[326] Chhavi Sharma et al. "SemEval-2020 Task 8: Memotion Analysis-the Visuo-Lingual Metaphor!" In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. 2020, pp. 759–773.

[327] Sara Shatford. "Analyzing the Subject of a Picture: A Theoretical Approach." In: (1986). DOI: 10.1300/j104v06n03_04.

[328] Ali Shirali and Moritz Hardt. "What Makes ImageNet Look Unlike LAION." In: *CoRR* abs/2306.15769 (2023). DOI: 10.48550/arXiv.2306.15769. arXiv: 2306.15769.

[329] Connor Shorten and Taghi M. Khoshgoftaar. "A survey on Image Data Augmentation for Deep Learning." In: *J. Big Data* 6 (2019), p. 60. DOI: 10.1186/s40537-019-0197-0.

[330] K Simonyan, A Vedaldi, and A Zisserman. "Deep inside convolutional networks: visualising image classification models and saliency maps." In: *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR. 2014.

[331] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015.

[332] Arnold WM Smeulders et al. "Content-based image retrieval at the end of the early years." In: *IEEE Transactions on pattern analysis and machine intelligence* 22.12 (2000), pp. 1349–1380. ISSN: 01628828. DOI: 10.1109/34.895972.

[333] T Smits and WJ Faber. *Chronic (classified historical newspaper images)*. Ed. by KB Lab. https://lab.kb.nl/dataset/chronic. 2018.

[334] Thomas Smits. *The Visual Digital Turn – Computer Vision and the Humanities*. video recording. KBR (Royal Library of Belgium), 2022.

[335] Thomas Smits and Melvin Wevers. "The agency of computer vision models as optical instruments." en. In: *Vis. commun.* 21.2 (2022), pp. 329–349.

[336] Francesco Solera, Simone Calderara, and Rita Cucchiara. "From Groups to Leaders and Back." en. In: *Group and Crowd Behavior for Computer Vision*. Elsevier, 2017, pp. 161–182. DOI: 10.1016/b978-0-12-809276-7.00010-2.

[337] Sebastian Stabinger and Antonio Rodriguez-Sanchez. "Evaluation of deep learning on an abstract image classification dataset." In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017, pp. 2767–2772.

[338] Matteo Stefanini et al. "Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain." In: *Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20*. Springer. 2019, pp. 729–740.

[339] Andreas Steiner et al. "How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers." In: *Trans. Mach. Learn. Res.* 2022 (2022).

[340] Mark Steyvers. "Combining feature norms and text data with topic models." In: *Acta Psychologica* 133.3 (2010), pp. 234–243.

[341] David G Stork. "Computer vision and computer graphics analysis of paintings and drawings: An introduction to the literature." In: *Computer Analysis of Images and Patterns: 13th International Conference, CAIP 2009, Münster, Germany, September 2-4, 2009. Proceedings 13*. Springer. 2009, pp. 9–24.

[342] Lise Stork et al. "Large-scale zero-shot learning in the wild: Classifying zoological illustrations." In: *Ecological informatics* 62 (2021), p. 101222.

[343] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. "Yago: a core of semantic knowledge." In: *Proceedings of the 16th international conference on World Wide Web*. 2007, pp. 697–706.

[344] Mohammed Suhail and Leonid Sigal. "Mixture-Kernel Graph Attention Network for Situation Recognition." In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): Ieee, 2019, pp. 10362–10371. DOI: 10.1109/iccv.2019.01046.

[345] Chen Sun et al. "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era." en. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: Ieee, 2017, pp. 843–852. DOI: 10.1109/iccv.2017.97.

[346] Qianru Sun, Bernt Schiele, and Mario Fritz. "A Domain Based Approach to Social Relation Recognition." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: Ieee, 2017, pp. 3481–3490. DOI: 10.1109/cvpr.2017.54.

[347] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks." In: *Advances in neural information processing systems* 27 (2014).

[348] Christian Szegedy et al. "Rethinking the Inception Architecture for Computer Vision." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2818–2826. DOI: 10.1109/cvpr.2016.308.

[349] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.

[350] Wei Ren Tan et al. "Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification." In: *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*. Ieee, 2016, pp. 3703–3707. DOI: 10.1109/icip.2016.7533051.

[351] Mark B Tappan. "Mediated moralities: Sociocultural approaches to moral development." In: *Handbook of moral development*. Psychology Press, 2006, pp. 369–392.

[352] Amara Tariq and Hassan Foroosh. "Learning semantics for image annotation." In: *arXiv preprint arXiv:1705.05102* (2017).

[353] The American Psychological Association. *Cultural Bias*. In: *The APA Dictionary of Psychology*. 2th. Houghton Mifflin Harcourt Publishing Company, 2015.

[354] Christopher Thomas and Adriana Kovashka. "Predicting the Politics of an Image Using Webly Supervised Data." In: *Advances in neural information processing systems*. Vol. 32. Curran Associates, Inc., 2019. DOI: 10.48550/arxiv.1911.00147.

[355] Christopher Thomas and Adriana Kovashka. "Predicting Visual Political Bias Using Webly Supervised Data and an Auxiliary Task." en. In: *International Journal of Computer Vision* 129.11 (2021), pp. 2978–3003. ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-021-01506-3.

[356] Christopher Thomas and Adriana Kovashka. "Preserving Semantic Neighborhoods for Robust Cross-Modal Retrieval." en. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 317–335. DOI: 10.1007/978-3-030-58523-5_19.

[357] Ilaria Tiddi and Stefan Schlobach. "Knowledge graphs as tools for explainable machine learning: A survey." In: *Artificial Intelligence* 302 (2022), p. 103627.

[358] Antoine Toisoul et al. "Estimation of Continuous Valence and Arousal Levels from Faces in Naturalistic Conditions." In: *Nature Machine Intelligence* 3.1 (2021), pp. 42–50. ISSN: 2522-5839. DOI: 10.1038/s42256-020-00280-0.

[359] Tiago Timponi Torrent et al. "Representing context in framenet: A multidimensional, multimodal approach." In: *Frontiers in Psychology* 13 (2022), p. 573.

[360] Paola Tubaro, Antonio A Casilli, and Marion Coville. "The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence." In: *Big Data &Society* 7.1 (2020), p. 2053951720919776.

[361] Miroslav Vacura et al. "Describing low-level image features using the comm ontology." In: *2008 15th IEEE International Conference on Image Processing*. Ieee. 2008, pp. 49–52.

[362] Lucia Vadicamo et al. "Cross-Media Learning for Image Sentiment Analysis in the Wild." In: *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. Venice: Ieee, 2017, pp. 308–317. DOI: 10.1109/iccvw.2017.45.

[363] Nicolò Oreste Pinciroli Vago et al. "Comparing CAM Algorithms for the Identification of Salient Image Features in Iconography Artwork Analysis." In: *J. Imaging* 7.7 (2021), p. 106. DOI: 10.3390/jimaging7070106.

[364] Pieter Vanneste et al. "Computer Vision and Human Behaviour, Emotion and Cognition Detection: A Use Case on Student Engagement." en. In: *Mathematics* 9.3 (2021), p. 287. DOI: 10.3390/math9030287.

[365] Elizabeth B. Varghese and Sabu M. Thampi. "A Deep Learning Approach to Predict Crowd Behavior Based on Emotion." en. In: *Smart Multimedia*. Ed. by Anup Basu and Stefano Berretti. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 296–307. DOI: 10.1007/978-3-030-04375-9_25.

[366] Ben Vassar. "The eclectic iconography of hyperpop." In: *The Michigan Daily* (2020).

[367] Ashish Vaswani et al. "Attention is All you Need." In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon et al. 2017, pp. 5998–6008.

[368] Emmeke Anna Veltmeijer, Charlotte Gerritsen, and Koen Hindriks. "Automatic emotion recognition for groups: a review." In: *IEEE Transactions on Affective Computing* (2021), pp. 1–1. ISSN: 1949-3045. DOI: 10.1109/taffc.2021.3065726.

[369] Gabriella Vigliocco et al. "The Neural Representation of Abstract Words: The Role of Emotion." In: *Cerebral Cortex* 24.7 (2014), pp. 1767–1777. ISSN: 1047-3211. DOI: 10.1093/cercor/bht025.

[370] Gabriella Vigliocco et al. "The representation of abstract words: What matters? Reply to Paivio's (2013) comment on Kousta et al.(2011)." In: (2013).

[371] Caterina Villani et al. "Varieties of abstract concepts and their multiple dimensions." en. In: *Language and Cognition* 11.3 (2019), pp. 403–430. ISSN: 1866-9808, 1866-9859. DOI: 10.1017/langcog.2019.23.

[372] Giulia Vilone and Luca Longo. "Explainable Artificial Intelligence: a Systematic Review." In: *CoRR* abs/2006.00093 (2020). arXiv: 2006.00093.

[373] Terrance de Vries et al. "Does object recognition work for everyone?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 52–59.

[374] Gang Wang et al. "Seeing People in Social Context: Recognizing People and Social Relationships." In: *Computer Vision – ECCV 2010*. Ed. by Kostas Daniilidis, Petros Maragos, and Nikos Paragios. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2010, pp. 169–182. DOI: 10.1007/978-3-642-15555-0_13.

[375] Hanchen Wang et al. "Scientific discovery in the age of artificial intelligence." In: *Nature* 620.7972 (2023), pp. 47–60.

[376] Limin Wang et al. "Better Exploiting OS-CNNs for Better Event Recognition in Images." In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. Santiago, Chile: Ieee, 2015, pp. 287–294. DOI: 10.1109/iccvw.2015.46.

[377] Weining Wang and Qianhua He. "A survey on emotional semantic image retrieval." In: *2008 15th IEEE International Conference on Image Processing*. 2008, pp. 117–120. DOI: 10.1109/icip.2008.4711705.

[378] Xiaoguang Wang et al. "Data modeling and evaluation of deep semantic annotation for cultural heritage images." In: *Journal of Documentation* (2021).

[379] Xiu-Shen Wei, Bin-Bin Gao, and Jianxin Wu. "Deep Spatial Pyramid Ensemble for Cultural Event Recognition." In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. Santiago, Chile: Ieee, 2015, pp. 280–286. DOI: 10.1109/iccvw.2015.45.

[380] Daniel S Weld and Gagan Bansal. "Intelligible artificial intelligence." In: *ArXiv e-prints, March 2018* (2018).

[381] Melvin Wevers. "Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990." In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 92–97. DOI: 10.18653/v1/W19-4712.

[382] Melvin Wevers and Thomas Smits. "The visual digital turn: Using neural networks to study historical images." In: *Digital Scholarship in the Humanities* 35.1 (2020), pp. 194–207.

[383] Maximilian Wich, Hala Al Kuwatly, and Georg Groh. "Investigating annotator bias with a graph-based approach." In: *Proceedings of the fourth workshop on online abuse and harms*. 2020, pp. 191–199.

[384] Maximilian Wich, Jan Bauer, and Georg Groh. "Impact of politically biased data on hate speech classification." In: *Proceedings of the fourth workshop on online abuse and harms*. 2020, pp. 54–64.

[385] Scott Workman, Richard Souvenir, and Nathan Jacobs. "Understanding and Mapping Natural Beauty." In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: Ieee, 2017, pp. 5590–5599. DOI: 10.1109/iccv.2017.596.

[386] Baoyuan Wu et al. "Tencent ML-Images: A Large-Scale Multi-Label Image Database for Visual Representation Learning." In: *IEEE Access* 7 (2019), pp. 172683–172693. ISSN: 2169-3536. DOI: 10.1109/access.2019.2956775.

[387] Wenzhuo Yang et al. "OmniXAI: A Library for Explainable AI." In: *arXiv* (2022). DOI: 10.48550/arxiv.2206.01612. eprint: 206.01612.

[388] Xingxu Yao et al. "Attention-Aware Polarity Sensitive Embedding for Affective Image Retrieval." In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): Ieee, 2019, pp. 1140–1150. DOI: 10.1109/iccv.2019.00123.

[389] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. "Situation Recognition: Visual Semantic Role Labeling for Image Understanding." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: Ieee, 2016, pp. 5534–5542. DOI: 10.1109/cvpr.2016.597.

[390] K. Ye and A. Kovashka. "ADVISE: Symbolism and External Knowledge for Decoding Advertisements." In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Vol. 11219 Lncs. Cham: Springer International Publishing, 2018, pp. 868–886. DOI: 10.1007/978-3-030-01267-0_51.

[391] Keren Ye et al. "Interpreting the Rhetoric of Visual Advertisements." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.4 (2019), pp. 1308–1323. ISSN: 1939-3539. DOI: 10.1109/tpami.2019.2947440.

[392] Eiling Yee. "Abstraction and concepts: when, how, where, what and why?" In: *Language, Cognition and Neuroscience* 34.10 (2019), pp. 1257–1265. ISSN: 2327-3798. DOI: 10.1080/23273798.2019.1660797.

[393] Eiling Yee and Sharon L Thompson-Schill. "Putting concepts into context." In: *Psychonomic bulletin and Review* 23 (2016), pp. 1015–1027.

[394] Wenchi Yeh and Lawrence W Barsalou. "The situated nature of concepts." In: *The American journal of psychology* 119.3 (2006), pp. 349–384.

[395] Yuanjun Xiong et al. "Recognize Complex Events from Static Images by Fusing Deep Channels." In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: Ieee, 2015, pp. 1600–1609. DOI: 10.1109/cvpr.2015.7298768.

[396] Sarah Zanette et al. "Automated decoding of facial expressions reveals marked differences in children when telling antisocial versus prosocial lies." en. In: *Journal of Experimental Child Psychology* 150 (2016), pp. 165–179. ISSN: 0022-0965. DOI: 10.1016/j.jecp.2016.05.007.

[397] Mehdi Zemni et al. "OCTET: Object-aware Counterfactual Explanations." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15062–15071.

[398] Xiaohua Zhai et al. "Scaling Vision Transformers." In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. Ieee, 2022, pp. 1204–1213. DOI: 10.1109/cvpr52688.2022.01179.

[399] C. Zhang. *Is Glitchcore a TikTok Aesthetic, a New Microgenre, or the Latest Iteration of Glitch Art?* Pitchfork. 2020.

[400] Dehai Zhang et al. "Knowledge Graph-Based Image Classification Refinement." In: *IEEE Access* 7 (2019), pp. 57678–57690. DOI: 10.1109/access.2019.2912627.

[401] Dengsheng Zhang, Md. Monirul Islam, and Guojun Lu. "A review on automatic image annotation techniques." en. In: *Pattern Recognition* 45.1 (2012), pp. 346–362. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2011.05.013.

[402] Zhanpeng Zhang et al. "From Facial Expression Recognition to Interpersonal Relation Prediction." en. In: *International Journal of Computer Vision* 126.5 (2018), pp. 550–569. ISSN: 1573-1405. DOI: 10.1007/s11263-017-1055-1.

[403] Zhanpeng Zhang et al. "Learning Social Relation Traits from Face Images." In: *2015 IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: Ieee, 2015, pp. 3631–3639. DOI: 10.1109/iccv.2015.414.

[404] Sicheng Zhao et al. "Affective Image Content Analysis: A Comprehensive Survey." en. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 5534–5541. DOI: 10.24963/ijcai.2018/780.

[405] Sicheng Zhao et al. "Computational emotion analysis from images: Recent advances and future directions." In: *Human Perception of Visual Information: Psychological and Computational Perspectives* (2022), pp. 85–113.

[406] Bolei Zhou et al. "Learning Deep Features for Discriminative Localization." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2921–2929. DOI: 10.1109/cvpr.2016.319.

[407] Xiangru Zhu et al. "Multi-Modal Knowledge Graph Construction and Application: A Survey." In: *IEEE Transactions on Knowledge and Data Engineering* (2022), pp. 1–20. DOI: 10.1109/tkde.2022.3224228.

[408] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. "Reasoning about object affordances in a knowledge base representation." In: *European conference on computer vision*. Springer. 2014, pp. 408–424.

[409] Mathias Zinnen et al. "Odor: The icpr2022 odeuropa challenge on olfactory object recognition." In: *2022 26th International Conference on Pattern Recognition (ICPR)*. Ieee. 2022, pp. 4989–4994.

[410] Mathias Zinnen et al. "Transfer Learning for Olfactory Object Detection." In: *arXiv preprint arXiv:2301.09906* (2023).

Appendix

Class-level AC Image Classification Performances

This section provides detailed tabular summary of the classification performance metrics for various models across the target AC clusters. All the methods have been tested with the same dataset splits and with the same evaluation metrics. We first share a set of tables that display classification metrics for deep learning models (VGG-16, ResNet-50, and ViT), then the classification performance of machine learning models (Decision Tree, Random Forest, XGB, SVM, Bayesian Network, Naive Bayes) and then performance metrics for SPK, including for hybrid methods that combine Knowledge Graph Embeddings (KGE) and Vision Transformers (ViT) embeddings.

(a) VGG-16

	Precision	Recall	F1-score	Support
comfort	0.54	0.86	0.66	603
danger	0.42	0.08	0.14	170
death	0.39	0.42	0.40	257
fitness	0.33	0.01	0.02	102
freedom	0.14	0.02	0.03	51
power	0.30	0.24	0.26	276
safety	0.25	0.03	0.05	33
accuracy			**0.47**	1492
macro avg	0.34	0.24	**0.23**	1492
weighted avg	0.42	0.47	0.41	1492

(b) ResNet-50

	Precision	Recall	F1-score	Support
comfort	0.54	0.83	0.66	603
danger	0.29	0.11	0.16	170
death	0.42	0.49	0.45	257
fitness	0.43	0.03	0.06	102
freedom	0.00	0.00	0.00	51
power	0.39	0.26	0.31	276
safety	0.14	0.03	0.05	33
accuracy			**0.48**	1492
macro avg	0.32	0.25	**0.24**	1492
weighted avg	0.43	0.48	0.42	1492

(c) ViT

	Precision	Recall	F1-score	Support
Comfort	0.58	0.84	0.68	603
Danger	0.47	0.22	0.30	170
Death	0.43	0.47	0.45	257
Fitness	0.57	0.04	0.07	102
Freedom	0.25	0.06	0.10	51
Power	0.36	0.29	0.32	276
Safety	0.36	0.12	0.18	33
Accuracy			**0.51**	1492
Macro Avg	0.43	0.29	**0.30**	1492
Weighted Avg	0.48	0.51	0.46	1492

Table V.1.2: Classification metrics for different DL models on ARTstract dataset

(a) Decision Tree

	Precision	Recall	F1-score	Support
comfort	0.53	0.57	0.55	605
danger	0.10	0.08	0.09	130
death	0.27	0.33	0.30	258
fitness	0.12	0.08	0.10	112
freedom	0.12	0.10	0.11	52
power	0.22	0.21	0.22	299
safety	0.06	0.06	0.06	36
Accuracy			**0.35**	1492
Macro avg	0.20	0.20	**0.20**	1492
Weighted avg	0.33	0.35	0.34	1492

(b) Random Forest

	Precision	Recall	F1-score	Support
comfort	0.52	0.82	0.64	605
danger	0.23	0.02	0.04	130
death	0.34	0.37	0.35	258
fitness	0.17	0.03	0.05	112
freedom	0.00	0.00	0.00	52
power	0.28	0.21	0.24	299
safety	0.33	0.03	0.05	36
Accuracy			**0.44**	1492
Macro avg	0.27	0.21	**0.20**	1492
Weighted avg	0.37	0.44	0.38	1492

(c) XGB

	Precision	Recall	F1-score	Support
comfort	0.53	0.80	0.63	605
danger	0.33	0.05	0.09	130
death	0.37	0.41	0.39	258
fitness	0.50	0.02	0.03	112
freedom	0.00	0.00	0.00	52
power	0.27	0.23	0.25	299
safety	0.00	0.00	0.00	36
Accuracy			**0.45**	1492
Macro avg	0.29	0.22	**0.20**	1492
Weighted avg	0.40	0.45	0.39	1492

Table V.1.3: Performance metrics for Decision Tree, Random Forest, and XGB.

(a) SVM

	Precision	Recall	F1-score	Support
comfort	0.53	0.81	0.64	605
danger	0.23	0.02	0.04	130
death	0.36	0.42	0.39	258
fitness	0.12	0.01	0.02	112
freedom	0.25	0.02	0.04	52
power	0.28	0.21	0.24	299
safety	0.33	0.03	0.05	36
Accuracy			**0.45**	1492
Macro avg	0.30	0.22	**0.20**	1492
Weighted avg	0.38	0.45	0.38	1492

(b) Bayesian Network

	Precision	Recall	F1-score	Support
comfort	0.52	0.75	0.62	605
power	0.12	0.05	0.07	130
safety	0.34	0.36	0.35	258
danger	0.22	0.09	0.13	112
fitness	0.00	0.00	0.00	52
freedom	0.27	0.19	0.22	299
death	0.00	0.00	0.00	36
Accuracy			**0.42**	1492
Macro avg	0.21	0.21	**0.20**	1492
Weighted avg	0.35	0.42	0.37	1492

(c) Naive Bayes

	Precision	Recall	F1-score	Support
comfort	0.57	0.72	0.64	605
danger	0.31	0.08	0.13	130
death	0.33	0.46	0.38	258
fitness	0.17	0.04	0.06	112
freedom	0.09	0.08	0.08	52
power	0.33	0.27	0.30	299
safety	0.09	0.06	0.07	36
Accuracy			**0.44**	1492
Macro avg	0.27	0.24	**0.24**	1492
Weighted avg	0.40	0.44	0.40	1492

Table V.1.4: Performance metrics for SVM, Bayesian Network, and Naive Bayes.

(a) KGE only (TransE trained for 1000 epochs)

	Precision	Recall	F1-score	Support
0	0.54	0.83	0.66	603
1	0.34	0.20	0.25	170
2	0.33	0.40	0.36	257
3	0.33	0.01	0.02	102
4	0.00	0.00	0.00	51
5	0.29	0.17	0.21	276
6	0.33	0.03	0.06	33
Accuracy			**0.46**	1492
Macro avg	0.31	0.23	**0.22**	1492
Weighted avg	0.40	0.46	0.40	1492

(b) Relative Representation of KGE (rel-KGE)

	Precision	Recall	F1-score	Support
0	0.57	0.83	0.67	603
1	0.44	0.22	0.30	170
2	0.40	0.41	0.40	257
3	0.27	0.03	0.05	102
4	0.25	0.04	0.07	51
5	0.30	0.25	0.27	276
6	0.22	0.06	0.10	33
Accuracy			**0.48**	1492
Macro avg	0.35	0.26	**0.27**	1492
Weighted avg	0.43	0.48	0.43	1492

(c) Relative Representation of ViT (relVit)

	Precision	Recall	F1-score	Support
0	0.57	0.84	0.68	603
1	0.35	0.21	0.26	170
2	0.43	0.49	0.46	257
3	0.44	0.04	0.07	102
4	0.28	0.20	0.23	51
5	0.39	0.21	0.28	276
6	0.00	0.00	0.00	33
Accuracy			**0.50**	1492
Macro avg	0.35	0.28	**0.28**	1492
Weighted avg	0.46	0.50	0.45	1492

Table V.1.5: Performance metrics for KGE, relKGE and relViT.

(a) Concatenation of absViT and absKGE (absViT || absKGE)

	Precision	Recall	F1-score	Support
0	0.61	0.75	0.67	603
1	0.42	0.21	0.28	170
2	0.42	0.42	0.42	257
3	0.21	0.10	0.13	102
4	0.22	0.16	0.18	51
5	0.33	0.36	0.35	276
6	0.19	0.09	0.12	33
Accuracy			**0.48**	1492
Macro avg	0.34	0.30	**0.31**	1492
Weighted avg	0.45	0.48	0.46	1492

(b) Concatenation of relViT and relKGE (relViT || relKGE)

	Precision	Recall	F1-score	Support
0	0.60	0.79	0.69	603
1	0.37	0.20	0.26	170
2	0.41	0.46	0.43	257
3	0.28	0.12	0.17	102
4	0.23	0.12	0.16	51
5	0.41	0.36	0.38	276
6	0.28	0.15	0.20	33
Accuracy			**0.50**	1492
Macro avg	0.37	0.31	**0.33**	1492
Weighted avg	0.47	0.50	0.47	1492

(c) Hadamard Product of relViT and relKGE (relViT $\odot$ relKGE)

	Precision	Recall	F1-score	Support
0	0.59	0.76	0.67	603
1	0.28	0.19	0.23	170
2	0.41	0.46	0.43	257
3	0.35	0.08	0.13	102
4	0.19	0.10	0.13	51
5	0.34	0.31	0.32	276
6	0.25	0.09	0.13	33
Accuracy			**0.48**	1492
Macro avg	0.34	0.28	**0.29**	1492
Weighted avg	0.44	0.48	0.45	1492

Table V.1.6: Performance metrics for hybrid methods combining KGE and ViT embeddings.

SPARQL Results for SituAnnotate Evaluation

To assess the SituAnnotate ontology's ability to answer key competency questions (CQs), a set of SPARQL queries was developed to showcase the ontology's proficiency in addressing these questions. This pairing between CQs and SPARQL queries was essential in evaluating how SituAnnotate can enable users to extract pertinent insights and contextually relevant data from annotation records. The section provides a Table (below) with the pairings and the test results. It exemplifies the ontology's role in facilitating context-aware explanations and insights, further underlining its significance in the domain of annotation data management and retrieval.

SituAnnotate Competency Questions and Corresponding SPARQL Queries

CQ	Competency Question	SPARQL Query	Pass

CQ1	Which countries have been the location of annotation situations, how many annotation situations were located in each country, and which country has been the location for the highest number of annotation situations?	`SELECT ?Country (COUNT(?AnnotationSituation) AS ?count)`
		`WHERE {`
		`  ?AnnotationSituation :atPlace ?Country .`
		`}`
		`GROUP BY ?Country`
		`ORDER BY DESC(?count)`
			Y

CQ2	Between the years 2020 and 2024, in which annotation situations has the image with ID *ARTstract_14978* been involved?	`SELECT ?AnnotationSituation ?Date`
		`WHERE {`
		`  :ARTstract_14978 :isInvolvedInAnnotationSituation ?AnnotationSituation .`
		`  ?AnnotationSituation :onDate ?Date .`
		`  FILTER(YEAR(?date) >= 2020 && YEAR(?date) <= 2024)`
		`}`
			Y

CQ3	What remuneration schemes have been used in annotation situations involving the *ARTstract* dataset?	`SELECT ?RemunerationScheme`
		`WHERE {`
		`  ?AnnotationSituation rdf:type :AnnotationSituation ;`
		`  :involvesDataset :ARTstract .`
		`  ?AnnotationSituation :involvesRemunerationScheme ?RemunerationScheme .`
		`}`
			Y

CQ4	What types of entities have been annotated?	
		`SELECT DISTINCT ?EntityType`
		`WHERE {`
		`  ?Annotation :aboutAnnotatedEntity ?Entity .`
		`  ?Entity a ?EntityType .`
		`}`
			Y

CQ5	Which images have been annotated using the lexical entry ”surfboard,” and what role did these annotations serve?	
		`SELECT ?Image ?annotationRole`
		`WHERE {`
		`  ?Annotation :aboutAnnotatedEntity ?Image .`
		`  ?Annotation :annotationWithLexicalEntry :le_surfboard .`
		`  ?Annotation :isClassifiedBy ?AnnotationRole .`
		`}`
			Y

CQ6	For the specific situation in which ”surfboard” was used to annotate the image with ID *ARTstract_14978*, what contextual factors were associated with the annotation situation?	
		`SELECT ?Country ?Date ?Dataset ?RemunerationScheme ?DetectionThreshold ?Annotator ?`
		`  PretrainDataset ?ModelArchitecture`
		`WHERE {`
		`  ?Annotation :aboutAnnotatedEntity :ARTstract_14978 .`
		`  ?Annotation :annotationWithLexicalEntry :le_surfboard .`
		`  ?AnnotationSituation :involvesAnnotation ?Annotation .`
		`  OPTIONAL {`
		`  ?AnnotationSituation :atPlace ?Country .`
		`  ?AnnotationSituation :onDate ?Date .`
		`  ?AnnotationSituation :involvesDataset ?Dataset .`
		`  ?AnnotationSituation :hasDetectionThreshold ?DetectionThreshold .`
		`  ?AnnotationSituation :involvesAnnotator ?Annotator .`
		`  ?Annotator :pretrainedOnDataset ?PretrainDataset .`
		`  ?Annotator :hasModelArchitecture ?ModelArchitecture .`
		`  ?AnnotationSituation :involvesRemunerationScheme ?RemunerationScheme .`
		`  }`
		`}`
			Y

CQ7	Which images have annotations classified under the role of "detected emotion" with an annotation strength exceeding 0.85, and what labels have been assigned to them?	
		`SELECT ?Image ?Label`
		`WHERE {`
		`  ?Image a :Image .`
		`  ?Annotation :aboutAnnotatedEntity ?Image ;`
		`  :isClassifiedBy :detected_emotion ;`
		`  :hasAnnotationStrength ?AnnotationStrength ;`
		`  :annotationWithLexicalEntry ?LE .`
		`  ?LE rdfs:label ?Label .`
		`  FILTER (?AnnotationStrength > 0.85)`
		`}`
			Y

CQ8	What concepts type annotations about the image with ID *ARTstract_14978*?	
		`SELECT ?Concept ?AnnotationRole ?AnnotationStrength`
		`WHERE {`
		`  ?Annotation :aboutAnnotatedEntity :ARTstract_14978 .`
		`  ?Annotation :isClassifiedBy ?AnnotationRole .`
		`  ?Annotation :hasAnnotationStrength ?AnnotationStrength .`
		`  ?Annotation :typedByConcept ?Concept .`
		`}`
			Y

CQ9	For each lexical entry (label) that the image with ID *ARTstract_14978* was annotated with, who was the Annotator that assigned that label?	
		`SELECT ?string ?Annotator`
		`WHERE {`
		`  :ARTstract_14978 :isInvolvedInAnnotationSituation ?AnnotationSituation .`
		`  ?AnnotationSituation :involvesAnnotation ?Annotation .`
		`  ?AnnotationSituation :involvesAnnotator ?Annotator .`
		`  ?Annotation :aboutAnnotatedEntity :ARTstract_14978.`
		`  ?Annotation :annotationWithLexicalEntry ?LexicalEntry .`
		`  ?LexicalEntry rdfs:label ?string .`
		`}`
			Y

CQ10	What types of annotations about the image with ID *ARTstract_14978* were all done by artificial annotators with the *visual transformer* model architecture?	
		`SELECT ?AnnotationClass ?Annotator ?Dataset`
		`WHERE {`
		`  ?AnnotationSituation :involvesAnnotation ?Annotation .`
		`  ?AnnotationSituation :involvesAnnotator ?Annotator .`
		`  ?Annotator :hasModelArchitecture :visual_transformer .`
		`  ?Annotator :pretrainedOnDataset ?Dataset .`
		`  ?Annotation :aboutAnnotatedEntity :ARTstract_14978 .`
		`  ?Annotation a ?AnnotationClass .`
		`  FILTER NOT EXISTS {`
		`  ?subClass rdfs:subClassOf ?AnnotationClass .`
		`  ?Annotation rdf:type ?subClass.`
		`  FILTER (?subClass != ?AnnotationClass)`
		`  }`
		`}`
			Y

CQ11	What are the caption annotations for the image with ID *ARTstract_14978*, and who are the annotators responsible for each caption annotation?	
		`SELECT ?Caption ?Annotator`
		`WHERE {`
		`  ?Annotation :aboutAnnotatedEntity :ARTstract_14978 .`
		`  ?Annotation a :ImageCaptionAnnotation .`
		`  ?Annotation rdfs:comment ?Caption .`
		`  ?AnnotationSituation :involvesAnnotation ?Annotation .`
		`  ?AnnotationSituation :involvesAnnotator ?Annotator .`
		`}`
			Y

Initial Experiments with LLMs

Large Language Models (LLMs) have traditionally been associated with natural language understanding and generation. However, there is emerging potential to leverage these models for high-level visual understanding tasks, such as AC image classification. This experiment explores the use of LLMs with our results to test performance and interpretability of AC image classification.

V.1.5.4 Approach

In these initial experiments, we aimed to utilize LLMs for AC image classification, relying on the image captions generated in Chapter III.2. Specifically, we employed the BLIP Instruct model from HuggingFace[1], using Vicuna-7b as the language model. The InstructBLIP model was introduced in [94], and we applied it with the following prompt:

> **PROMPT:** *What abstract concept among ['comfort', 'danger', 'death', 'fitness', 'freedom', 'power', 'safety'] does this image depict?*

V.1.5.5 Results

The results, as shown in the Table below, indicate that the model struggled to follow the instructions for multi-class classification. Instead, it created numerous different classes, leading to metrics of zero accuracy for each class. The output classes included variations of the expected classes, making evaluation challenging.

V.1.5.6 Discussion

The use of the 7B parameter model may not be sufficient to generalize to new tasks, and few-shot prompting was not employed. It appears that models with higher capacities, such as 30-50B, may be necessary for better generalization when using in-context prompting. Additionally, the model quantization to 4 bits might have slightly impacted accuracy, but experimental evidence suggests that the loss during the inference task is not significant. The current approach to multimodal tasks, while effective for concrete and perceptually-bound concepts, may not be suitable for high-level semantic understanding. To tackle tasks requiring advanced reasoning, LLMs are limited because they lack the ability to reason effectively.

[1] https://huggingface.co/Salesforce/instructblip-vicuna-7b. Access date: October 2023.

Class	Precision	Recall	F1-Score	Support
”comfort”	0.00	0.00	0.00	0
”danger”	0.00	0.00	0.00	0
”danger”	0.00	0.00	0.00	0
”death”	0.00	0.00	0.00	0
15th century the image depicts an 80-foot long ship with a	0.00	0.00	0.00	0
16th century depiction of a man in fur hats	0.00	0.00	0.00	0
17	0.00	0.00	0.00	0
1762 painting the abstract concept depicted in the 1548 engra	0.00	0.00	0.00	0
1860s sailor bartending	0.00	0.00	0.00	0
18th birthday	0.00	0.00	0.00	0
1903 ship wreck	0.00	0.00	0.00	0
1920s-30	0.00	0.00	0.00	0
1950s comic book cover	0.00	0.00	0.00	0
1984	0.00	0.00	0.00	0
19th century dance	0.00	0.00	0.00	0
1: ’freedom’	0.00	0.00	0.00	0
1st or second born child	0.00	0.00	0.00	0
2	0.00	0.00	0.00	0
mountain which among `['comfort', 'danger']`	0.00	0.00	0.00	0
mountain climbing	0.00	0.00	0.00	0
mountains	0.00	0.00	0.00	0
nutrition	0.00	0.00	0.00	0
power	0.22	0.23	0.23	276
protection	0.00	0.00	0.00	0
racy	0.00	0.00	0.00	0
risk	0.00	0.00	0.00	0
safety	0.00	0.00	0.00	33
sexual	0.00	0.00	0.00	0
sexual freedom	0.00	0.00	0.00	0
sexuality	0.00	0.00	0.00	0
ship	0.00	0.00	0.00	0
space	0.00	0.00	0.00	0
strength	0.00	0.00	0.00	0
survive	0.00	0.00	0.00	0
the ”finger”	0.00	0.00	0.00	0
☠️ danger	0.00	0.00	0.00	0
☯️ danger	0.00	0.00	0.00	0
freedom	0.00	0.00	0.00	0
🌞 freedom	0.00	0.00	0.00	0

Class	Precision	Recall	F1-Score	Support
the power of freedom	0.00	0.00	0.00	0
the power of love over danger, death and comfort	0.00	0.00	0.00	0
the power of sea	0.00	0.00	0.00	0
the presence	0.00	0.00	0.00	0
the word ”power” depicts this image among its five words	0.00	0.00	0.00	0
the word safety	0.00	0.00	0.00	0
transportation	0.00	0.00	0.00	0
water	0.00	0.00	0.00	0
©	0.00	0.00	0.00	0
€power	0.00	0.00	0.00	0
	0.00	0.00	0.00	0
💘	0.00	0.00	0.00	0
⚔️ danger	0.00	0.00	0.00	0
❄️ power	0.00	0.00	0.00	0
❌ danger	0.00	0.00	0.00	0
❌ power	0.00	0.00	0.00	0
❌ danger	0.00	0.00	0.00	0
❤️	0.00	0.00	0.00	0
(a warrior)	0.00	0.00	0.00	0
🗻	0.00	0.00	0.00	0
💧 water	0.00	0.00	0.00	0
✨ (make life happen)	0.00	0.00	0.00	0
👠 passion	0.00	0.00	0.00	0
🔴 freedom	0.00	0.00	0.00	0
power	0.00	0.00	0.00	0
🧢 comfort	0.00	0.00	0.00	0
Accuracy	0.12			1492
Macro Avg	0.00	0.00	0.00	1492
Weighted Avg	0.09	0.12	0.10	1492