



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN
COMPUTER SCIENCE AND ENGINEERING

Ciclo 38

Settore Concorsuale: 01/B1 - INFORMATICA

Settore Scientifico Disciplinare: INF/01 - INFORMATICA

DEEP LEARNING MODELS FOR PERSON RE-IDENTIFICATION AND MULTI-
PERSON TRACKING

Presentata da: Salvatore Fiorilla

Coordinatore Dottorato

Paola Salomoni

Supervisore

Andrea Asperti

Co-supervisore

Maurizio Gabbrielli

Esame finale anno 2026

*A complex system that works is invariably found to have
evolved from a simple system that worked.*

— John Gall

Declaration

I declare that, except where explicitly acknowledged, the content of this dissertation is my own original work and has not been submitted, in whole or in part, for any other degree or qualification at this or any other university. This dissertation reflects my personal research and does not include work done in collaboration with others, except where specifically stated in the text and acknowledgements.

October, 2025

Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisor, Professor Andrea Asperti, for guiding me through these years of work with his experience and knowledge, and for inspiring me every day.

I also thank Professor François Brémont for hosting me for six months in his research team at INRIA, for his patience and care during that time. It was an outstanding period of personal and professional growth that I will always remember.

Of course, genuine thanks go to my best friends and all the people with whom I shared beautiful and unforgettable moments of hard daily work. I am especially grateful to Fabio, Davide, and Aglind, who went from colleagues to friends.

Finally and importantly, I want to extend my deepest gratitude to my Family and Sofia, my other half, for encouraging and supporting me throughout this three-year journey.

Abstract

This thesis presents a systematic study of Person Re-Identification (ReID), a core computer vision task enabling the association and re-identification of individuals across different times and camera views. ReID enables applications from multi-camera tracking and video surveillance to non-invasive behavioral analysis, playing a key role in how automated systems balance technical innovation with privacy. The research is structured around four pillars: (RQ1) identification of challenges and trends in supervised and unsupervised ReID; (RQ2) exploration of embedding methods to extract identity-specific features while minimizing transient attributes like pose and background; (RQ3) evaluation of robustness to domain shifts and strategies for domain adaptation; and (RQ4) assessment of integration into multi-person tracking systems. The thesis combines theoretical analysis, empirical evaluation, and proof-of-concept implementations. Experiments with Denoising Diffusion Models demonstrate disentangling stable identity features from transient cues and potential for generative data augmentation. Domain adaptation studies show that carefully selected combinations of heterogeneous datasets improve cross-domain performance, highlighting the importance of data-centric adaptation strategies. Video-based ReID investigations show that Transformer architectures achieve state-of-the-art results by leveraging temporal consistency across frames. An ablation study on ReID integration in multi-person tracking reveals robustness depends on the deployment scenario: high benchmark performance does not ensure reliable real-world tracking. Overall, excelling on benchmarks alone is insufficient. The thesis

highlights the importance of integrating discriminative feature extraction, domain adaptation, and system-level design into a coherent framework. Future directions include unsupervised learning, generative augmentation for domain shifts, multimodal pre-training with temporal modeling, and continual adaptation. Ethical considerations are addressed, emphasizing the responsible deployment of ReID technologies in practice.

Keywords

Person Re-Identification, Computer Vision, Deep Learning, Tracking Systems

Contents

Abstract	iii
1 Introduction	1
1.1 Technical Background	2
1.1.1 Image-based Person Re-Identification	3
1.1.2 Video-based Person Re-Identification	5
1.1.3 Datasets	6
1.1.4 Metrics	12
1.2 Thesis outline	15
1.3 Research questions	16
2 Literature Review on Person Reidentification	19
2.1 Search and selection of papers	20
2.1.1 Datasets	20
2.2 Supervised Techniques	21
2.2.1 Feature Learning Approach	22
2.2.2 Metric Learning Approach	29
2.3 Unsupervised Techniques	33
2.3.1 Unsupervised Domain Adaptation methods	34
2.3.2 Fully unsupervised methods	39
2.4 Performance Analysis and Comparisons	45
2.4.1 Supervised ReID performance	45
2.4.2 Unsupervised ReID Performance	46
2.4.3 Method Comparisons	49

2.4.4	Conclusions	49
2.5	Conclusions on Research Questions	51
2.5.1	Insight on research question 1	52
2.5.2	Insight on research question 2	53
2.5.3	Insight on research question 3	54
3	PoC 1: Generation and Manipulation of embeddings	55
3.1	Denoising Diffusion Models	56
3.1.1	The denoising network	58
3.1.2	Sampling	59
3.2	The proposed Approach	60
3.2.1	Methodology	63
3.2.2	Neural Networks architectures	65
3.3	Evaluation	67
3.3.1	Latent space exploration	68
3.3.2	Ablation and alternatives	70
3.3.3	Conclusions	72
3.4	Conclusions on Research Questions	73
3.4.1	Insight on research question 2	73
3.4.2	Insight on research question 3	74
4	PoC 2: Domain Adaptation experiments	77
4.1	Technical Background	78
4.1.1	Models	79
4.1.2	Data Augmentation	85
4.2	Experiments results	87
4.2.1	Cross-Dataset Tests	87
4.2.2	Training on the Union of Datasets	89
4.2.3	Additional Experiments	91
4.2.4	Conclusions	94
4.3	Conclusions on Research Questions	96
4.3.1	Insight on research question 3	96

4.3.2	Insight on research question 4	96
5	At the intersection of Person ReID and Tracking	99
5.1	Image-based ReID evaluation	100
5.2	Video-based reid evalutation	104
5.3	MOT Tracking Benchmark evaluation	108
5.4	Conclusions and Discussions	112
5.5	Conclusions on Research Questions	112
5.5.1	Insight on research question 4	113
6	Conclusions	115
6.1	Scientific Conclusions	115
6.2	Ethical e Social Considerations	117

List of Tables

1.1	Person ReID Datasets	18
2.1	Market and Duke dataset information.	21
2.2	Supervised comparison in Market1501-Duke	46
2.3	Supervised comparison in Market1501-MSMT17	47
2.4	Unsupervised person re-identification papers	48
2.5	Top-5 Sup and Unsup Person Reid papers	50
4.1	CLIP-ReID framework evaluation	85
4.2	Cross-Dataset test (tr. on Market-1501)	88
4.3	Cross-Dataset test (tr. on DukeMTMC-reID)	88
4.4	Cross-Dataset test (tr. on Market1501+Airport)	89
4.5	Cross-Dataset test (tr. on Market1501+Airport, no REA)	90
4.6	Cross-Dataset test (tr. on Market1501+Airport+DukeMTMC-reID)	92
4.7	Cross-Dataset test (tr. on Market1501+Airport+ENTIRE-ID+CUHK03)	92
5.1	Models evaluation on Market1501	103
5.2	Models evaluation on iLIDS-Vid	107
5.3	MOT17 ablation study	110
5.4	MOT17 testing evaluation	111

List of Figures

1.1	Person Re-Identification Task	2
1.2	Reid data samples from different datasets	6
1.3	Market-1501 images samples	9
1.4	ENTIRE-ID dataset images samples	11
2.1	Vanilla model for supervised person ReID	22
2.2	SSP-ReID framework	25
2.3	ViT framework	27
2.4	EMA teacher-student framework	36
2.5	Vanilla model for unsupervised person ReID	40
2.6	Supervised and Unsupervised comparisons	51
3.1	Direct and Reverse diffusion process.	57
3.2	Conditional generative model and its inversion.	61
3.3	Conditional generation samples	62
3.4	Single and double network inversion	64
3.5	People reconstruction examples	65
3.6	Denosing Network Architecture	67
3.7	Main blocks of the Denoising Network.	68
3.8	Person Reidentification Ranklists.	69
3.9	Inherent clustering of latent space.	70
3.10	Closest and distance individuals.	71
3.11	Effect of latent variables on generation.	75

4.1	Clip image classification framework.	82
4.2	The baseline model.	83
4.3	The CLIP-ReID framework.	84
4.4	REA: Random erasing augmentation	86
5.1	Strong Baseline Netowork Architecture.	101
5.2	SEAS framework.	102
5.3	PSTA framework.	104
5.4	BoT-SORT Algorithm	109

Chapter 1

Introduction

Person re-identification has become a key enabler feature for a range of higher-level applications in computer vision. It allows systems to associate and re-identify individuals observed at different moments. Modern applications include tracking systems, smart video surveillance, and enhanced analytics in environments such as airports [1], [2], university campuses [3], [4], and public transport and public hubs [5].

The feature has the potential to enable future non-invasive applications, for example, systems could perform anonymous analysis in sports [6] or, as demonstrated in [7], in transit stations to collect Origin-Destination data in order to optimize and assess the impact of public transportation operations, without identifying individuals.

Thus, person re-identification not only represents a technical challenge within computer vision but also plays a central role in shaping how intelligent systems can operate responsibly in society, balancing technical innovation with privacy and ethical considerations.

The scientific contribution of this thesis is a systematic study of the Person Re-Identification task, starting from a critical analysis of the literature (Chapter 2), investigating the two main challenges of the field: identity-specific feature extraction (Chapter 3) and robustness to domain shifts (Chapter 4) - and culminating in the implementation of an integrative approach

applied to multi-person tracking scenarios (Chapter 5). The work is guided by four research questions, connecting foundational insights with practical applications and aiming to close the gap between fundamentals of Re-ID and multi-person tracking.

1.1 Technical Background

The Person Re-Identification (ReID or Person ReID) problem is a computer vision task aimed at matching images of the same individual captured from multiple views, typically by non-overlapping camera networks [8]–[11].

In standard inference protocols (references on Table 1.1), a Person ReID system receives a probe image of a person, referred to as the *query*, and compares it to a predefined set of images, referred to as *the gallery*. The system computes a *similarity measure* between the query and each gallery image, producing a ranked list of individuals ordered by their likelihood of matching the query, as illustrated below in Figure 1.1.



Figure 1.1: [10] The Person Re-Identification task: the query image of a woman is matched against the gallery set by n independent cameras. The output is an ordered list of gallery images that the model predicts depict the same person as the query.

The dataset splitting and evaluation process is originally adopted from image retrieval protocols ([1]), and the challenge of handling unseen identi-

ties closely aligns with the fundamentals of zero-shot learning ([12]). However, Person ReID introduces a more structured similarity matching, where individuals may appear similar in short-term scenarios but can change significantly in long-term ReID ([13]). A further unique challenge is the presence of unseen, potentially infinite identities at test time ([1]) also named as open-world settings and is inherited from real-world deployments where the system require to correctly match both: novel and seen identities. In the next sections, we will delve into the technical details of modern ReID systems. These systems have attracted increasing interest in both academia and industry, driven by advances in deep learning, particularly in computer vision, and by their applications in intelligent surveillance systems ([11]). Despite this growing interest, we will also see that Person ReID remains challenging due to the inherent difficulties of multi-camera environments, as well as frequent occlusions including variations in lighting, viewpoint, and image quality, a([8], [11]). To better illustrate these challenges, we distinguished between image-based (frame) and video-based approaches. Both rely on neural networks that extract low-dimensional feature representations for each individual and compute similarity scores.

Next section 1.1.1 discusses frame-based methods and challenges, while Section 1.1.2 focuses on video-based. Section 1.1.3 presents an overview of current benchmark datasets, and Section 1.1.4 describes the evaluation metrics used in person re-identification.

1.1.1 Image-based Person Re-Identification

At its core, image-based person re-identification focuses on neural networks that learn distinctive characteristics of individuals from images, allowing the system to discriminate among them ([14]). Such models generate a feature vector: a dense, low-dimensional representation that encodes the properties or semantics of the image and the person which is also called embedding ([3], [15]). Similar people should have similar embeddings, enabling deep learning models to process and compare them efficiently. Thus, a simi-

larity metric is computed usually with cosine similarity or euclidean distance to establish the score’s match ([16]–[18]).

Real-world scenarios presents variations in input images that can cause embeddings of the same person to differ: the network activation strongly depend on its input distribution. So, the network must learn to map very different appearance of the same individuals to the same cluster in feature space, minimizing matching errors ([16], [19]).

In this context, a good neural network remains stable and discriminative if it can effectively handle the following image diversity factors:

- **Viewpoint and pose variations:** the same person may appear in different angles and postures.
- **Lighting variations:** changes in illumination can significantly alter the appearance of clothing and body features.
- **Image quality variations:** low resolution, noise, or motion blur can hinder reliable feature extraction.
- **Cross-camera and scene variations:** individuals may appear across different cameras and diverse backgrounds.
- **Partial occlusions:** other people or objects may block parts of the body.
- **Clothing changes:** the same person may wear different outfits over time, or two different people may wear similar clothes, making classical appearance features less reliable and requiring more intrinsic, biometric-based cues [20].
- **Scale and size variations:** varying distances from the camera affect the apparent size of the person.

Despite these challenges, research has made significant progress, with most modern benchmarks reporting a Rank-1 score of at least 90% [9], [10],

[21]. This success has been driven by advances in deep learning architectures, particularly ResNet [19], [22] and Vision Transformers [21], [23], [24], along with the development of larger, higher-quality datasets [1], [3]–[5], [25]–[27]. We shall see them in Section 1.1.3.

1.1.2 Video-based Person Re-Identification

Conversely to frame-based Person Re-Identification, where a single embedding is computed at each frame, in video-based ReID the input is a *Tracklet*: a sequence of cropped images of a single individual, extracted from a video (sequence of frames) by a tracking algorithm. ([2], [28]). The way of addressing the match remains the same but the network is designed to produce a single embedding from the input tracklet, capturing both the person’s appearance and the temporal dynamics across frames. ([29]–[31])

In addition to the challenges presented in image-based tasks, there are specific problems that arise when moving to video-based ReID:

- **Temporal dynamics:** the person may change posture, walking style, or orientation along the sequence.
- **Motion blur and rapid variations:** Movement can introduce blur or sudden changes in appearance.
- **Frame selection/redundancy:** not all frames contain useful information; representative frames must be aggregated or filtered.
- **Temporal occlusions:** objects or other people may block the view in some frames but not in all.
- **Tracklet fragmentation:** a sequence of the same person may be interrupted or split due to tracking loss.
- **Inter-frame consistency:** object features must remain consistent across the sequence to create a reliable aggregated embedding.



Figure 1.2: Reid data samples from different datasets

- **Computational cost:** processing long sequences requires more resources than single images.

To have a better idea, current state-of-the-art showed an average map around 80% and Rank-1 around 85% such as [30], [32] on benchmarks that we will discuss better in the next section [2], [4], [33].

1.1.3 Datasets

In recent years, the Person ReID community has collected several large-scale datasets, primarily consisting of cropped pedestrian images, each representing a single individual. These bounding boxes may be hand-crafted or extracted automatically by object detection and/or tracking algorithms [34], [35]. Some samples are illustrated in Figure 1.2.

Each image is annotated with an identity label and the camera ID, serving as ground-truth information. Some datasets further include additional modalities, such as thermal imagery or sketches.

For model evaluation and training, datasets are typically divided into training and testing sets, ensuring that identities in the test set do not appear in training; the test sets consist of a *query set*, containing images as search inputs and a *gallery set* containing the database where the model retrieves images of the same individual.

Each query identity must appear in at least one (preferably more) gallery image captured by a different camera to enable cross-camera matching. During evaluation, the model searches the gallery for images depicting the same person as each query, following the standard benchmark protocol adopted in most ReID datasets [9], [10].

In this section, we will introduce the most widely employed datasets among the works discussed in this study, allowing for a fair comparison across various research contributions. We reported a summaries of their details in Table 1.1.

CUHK03

CUHK03 is a dataset released in 2014 by Li, Zhao, Xiao, *et al.* [25] at the Chinese University of Hong Kong. Compared to earlier datasets [36]–[41], CUHK03 exhibits a greater magnitude in terms of data volume with 13,164 bounding boxes and 1,360 identities extracted from footage recorded by 6 cameras, with each identity being captured by two cameras and having an average of 4.8 images per identity in each of the two camera views.

Market-1501

Introduced by Zheng, Shen, Tian, *et al.* [3] in 2015, Market-1501 is a person re-identification dataset consisting of 32,668 bounding boxes representing, as the name implies, 1,501 different identities (some examples can be seen in Figure 1.3). The dataset was created from footage recorded by six cameras in front of a campus market. The dataset also includes junk and distractor images. Market was also one of the first datasets where each identity could have multiple images under each camera, as opposed to the usual approach of having each identity only captured by 2 cameras [25], [41], [42], meaning that both the query and the gallery set may contain images from multiple cameras of the same identity.

MARS

The MARS (Motion Analysis and Re-identification Set) dataset, introduced by Zheng, Bie, Sun, *et al.* [33] in 2016, is a large-scale video-based person re-identification benchmark extended from the Market-1501 dataset. It contains 1,261 identities captured by 6 cameras, with a total of 20,715 tracklets and approximately 1,191,003 bounding boxes. The dataset includes both automatically generated tracklets using the Deformable Part Model (DPM) detector and Deep Matching tracker, as well as manually annotated sequences. Each identity has an average of 13.2 tracklets, with sequence lengths varying significantly. MARS presents realistic challenges including distractors (false detections), occlusions, viewpoint variations, and diverse motion patterns. The large scale and diversity of MARS have made it one of the most widely-used benchmarks for evaluating video-based person re-identification methods, particularly those exploiting temporal aggregation and motion modeling. *Note:* The dataset availability online is limited.

DukeMTMC-reID

DukeMTMC-reID (Duke **M**ulti **T**arget **M**ulti **C**amera re-identification) [4] is a dataset released by Ristani, Solera, Zou, *et al.* in 2016. It was crafted from footage recorded by 8 high quality cameras placed on Duke University’s campus. It is one of largest dataset consisting of hand-crafted bounding boxes [43] extracted from high quality images. The footage consists of an 85 minute recording taken from the time in between lectures (to guarantee high pedestrian activity) for each camera. The cameras record at 1080p and 60 frames per second. Each person was manually tracked by recording the person’s foot point of contact with the ground. Finally, for each of the eight cameras, the first five minutes of footage were reserved for training, and the remaining for testing. **Important:** The dataset was officially withdrawn by Duke University in June 2019 following investigations into its misuse for surveillance purposes and ethical concerns regarding consent and privacy. Despite its withdrawal, DukeMTMC-reID and DukeMTMC-VideoReID continues to

appear in academic literature. Researchers should be aware of these ethical considerations and the dataset’s unavailability for new research projects.

DukeMTMC-VideoReID

The DukeMTMC-VideoReID dataset, introduced by Ristani, Solera, Zou, *et al.* [4] in 2016, is a video-based person re-identification benchmark derived from the DukeMTMC (Duke Multi-Target, Multi-Camera Tracking) dataset. It contains 1,404 identities (702 for training and 702 for testing) captured by 8 high-resolution cameras deployed on Duke University’s campus. The dataset comprises 4,832 tracklets (2,196 for training and 2,636 for testing) with varying sequence lengths. Each tracklet contains manually annotated bounding boxes, ensuring high-quality person localization. The dataset captures challenging real-world scenarios, including occlusions, lighting variations, different viewpoints, and crowded scenes with multiple pedestrians.



Figure 1.3: Some of the Market-1501 images samples: the top row contains two distinct identities that share very similar appearance, the middle row consists of two distinct identities that are easily told apart, and the bottom row consists of distractor and junk images.

MSMT17

The MSMT17 (**M**ulti-**S**cene **M**ulti-**T**ime) person re-identification dataset, introduced in 2017 by Wei, Zhang, Gao, *et al.* [26], is one of the largest and most challenging supervised person ReID datasets. It was derived from 180 hours of video footage captured by 15 cameras installed on a university campus, including 12 cameras covering outdoor scenes under various weather conditions and 3 cameras for indoor environments.

The pedestrian bounding boxes were detected automatically using a Faster R-CNN [44], which improves on the DPM [45] object detector that was the standard in earlier works [3], [25] and provides fairly accurate detection. The datasets contains 126,441 bounding boxes spanning 4,101 identities, which is a significant improvement over earlier datasets, all of which contained less than 2,000 identities [3], [25], [41], [42], as can be seen in Table 1.1.

Airport

Airport [1] is a person re-identification dataset constructed from footage recorded inside a major airport. The footage was recorded over the course of 12 hours in a single day by six cameras placed after the airport’s security checkpoints. Each 12 hour video was randomly split into 40 five minute clips which were then used to create the dataset.

Airport’s main novelty lies in its camera network and its environment: the cameras used were the airport’s actual security cameras, whose complex orientation and low resolution introduces additional challenges. Furthermore, the inside of an airport contains a greater variety of people, and more varying crowd dynamics than a university campus.

The dataset was created during the deployment of a real-time intelligent surveillance system [46], with bounding boxes generated by the ACF framework [47] and tracking performed using the KLT tracker [48] and FAST corner features [49]. It contains 9,651 identities, of which 1,382 reappear across multiple cameras, while the remaining unpaired identities were included to increase the dataset’s complexity.



Figure 1.4: Examples of images from the ENTIRE-ID dataset.

ENTIRE-ID

ENTIRE-ID [5] is, to the best of our knowledge, the most recent and most extensive supervised person re-identification dataset, having been released in 2024 and containing more than four million images¹.

The ENTIRE-ID datasets contains 4.45 million images spanning 13,540 identities; some example are shown in Figure 1.4. The source footage was taken from 37 publicly available internet cameras located in four different continents. This massively improves the standard for dataset size: some of the closest supervised datasets in terms of size are the LaST dataset [50], created from movie footage and comprising 228,156 images spanning 10,862 identities, and the MARS dataset [33], which is a video extension of the Market-1501 dataset [3] containing more than one million of bounding boxes, but with the same identities and cameras of Market-1501.

Bounding boxes have been created by means of the YOLOv8 [51] object detection model, integrated with the ByteTrack [52] tracking algorithm to correctly identify people in consecutive video frames. The face of each person contained in the dataset is blurred, both to preserve the subject’s privacy as well as to push models away from learning and exploiting facial features.

¹As of the time of writing, only the test set has been released.

LUPerson

LUPerson [27] is an unlabeled person re-id dataset, containing more than 4 million images spanning over 200k identities, extracted from over 46k youtube videos. The videos were selected by searching for streetview footage recorded in 100 big cities, resulting in roughly 680 to 760 videos per city. The bounding boxes were extracted using YOLO-v5². LUPerson was conceived as an alternative to ImageNet [53] for unsupervised pre-training which, given the modest size of most supervised datasets, may play an important role in person re-id.

iLIDS-VID dataset

The iLIDS-VID dataset, introduced by Li, Zhu, and Gong [2] in 2014, is a video-based person re-identification benchmark derived from the i-LIDS Multiple-Camera Tracking Scenario (MCTS), captured at an airport arrival hall in the UK. It consists of 600 image sequences of 300 individuals, each recorded from two non-overlapping camera views. The sequence length varies from 23 to 192 frames, with an average of approximately 73 frames per tracklet. The dataset features challenging real-world conditions, including cluttered backgrounds, severe occlusions, and significant variations in illumination and viewpoint. Owing to these complexities, iLIDS-VID has become a widely used benchmark for evaluating video-based person re-identification methods that exploit temporal dynamics and motion patterns.

1.1.4 Metrics

In order to evaluate the person re-identification models, these are evaluated on its ability to retrieve, for each query image, images from the gallery depicting with cross-camera images of the same person. As in information retrieval systems that provide a ranked sequence of documents, person re-identification produces a ranked list of gallery images for each query. It

²<https://github.com/ultralytics/yolov5>

represents an ordered list of the probable found matches [9], [21], [54].

Two metrics are widely used for compute this evaluation: Mean Average Precision (*mAP*) and Cumulative matching characteristic (*CMC*) curve. ([9], [10], [14]) We will introduce them right below.

Mean Average precision (mAP)

In evaluating ReID systems of provide a good ranked sequence, it is important to consider the order in which returned document are presented. Modern datasets, such as Market-1501[3], often have multiple ground truths per query, requiring weighted metrics on the number of the queries.

The Mean Average Precision (mAP) metric is calculated as the mean of average precision values across all queries. Average precision considers the precision of the model at various ranks, with precision at rank k denoted as P_k (or $P@k$). Below, we recall its formal definition, for the sake of completeness.

Given a sorted scored list $G' = (g_1, \dots, g_n)$ of items, each labeled by a ground truth t_i , a query sample q and its ground truth t_q (in the case of person re-id the samples will be images and the ground truths will be the person's identity) we can define the precision at k as the ratio of true positives within the first k samples of G' and k . A sample g_i is considered a true positive if its ground truth t_i matches t_q . Thus the number of true positives within the first k samples for query q can be written as:

$$TP_k(q) = |\{g_i \in (g_1, \dots, g_k) : t_i = t_q\}| \quad (1.1)$$

The precision $P_k(q)$ at k for query sample q is:

$$P_k(q) = \frac{1}{k} TP_k(q) \quad (1.2)$$

We also need to define the relevance-at-k function:

$$rel_k(q) = \begin{cases} 1 & \text{if } t_k = t_q \\ 0 & \text{else} \end{cases} \quad (1.3)$$

Then the average precision $AP(q)$ for a given query q is:

$$AP(q) = \frac{1}{TP_n(q)} \sum_{k=1}^n P_k(q) \cdot rel_k(q) \quad (1.4)$$

So, mAP is simply the mean of the AP function over all query samples:

$$mAP = \frac{1}{|Q|} \sum_{q \in Q} AP(q) \quad (1.5)$$

Cumulative Matching Characteristic (CMC) curve

The *CMC* measures the fraction of queries (hence, the likelihood) whose correct match appears within the top- k results.[54], [55] Unlike *mAP*, it ignores the number of true matches and their ranks, checking only if at least one correct match is within the top- k . Common k values are 1, 5, 10, and 20 referred to as Rank-1, Rank-5, Rank-10, Rank-20 accuracies. Plotting *CMC* for all k up to the gallery length yields a monotonically increasing curve between 0 and 1, representing the cumulative matching probability across ranks[9], [21], [54].

The metric is often denoted CMC_k or *Rank- k* (or similar notations), where k represents the rank threshold. This metric, along with mean average precision, is widely used for person re-id evaluation [3]. Formally, given $D = (T, Q, G)$ (training, query, gallery sets respectively), a model M , and a distance function d we have that:

$$CMC_k(M, Q, G, d) = \frac{1}{|Q|} \sum_{q \in Q} acc_k(q, G') \quad (1.6)$$

where G' is the output of model M on G sorted according to the distance d from the output of model M on the query sample q (in layman's terms, G' will be the gallery set sorted according to how close the model currently thinks each gallery image is to query image q), and where acc_k is the accuracy at k function:

$$acc_k(x, S) = \begin{cases} 1 & \text{if a sample matching } x \text{ is in the top-}k \text{ samples of } S \\ 0 & \text{otherwise} \end{cases} \quad (1.7)$$

As mentioned, in person re-id "*a sample matching* x " means an image of the same person as x , though the definition easily extends to classification or other supervised retrieval tasks.

Consequently, **Rank-1 Accuracy** is widely adopted alongside *mAP* in benchmarks and it is formulated as:

$$\text{Rank-1 Accuracy (R1)} = \frac{1}{N} \sum_{i=1}^N \delta(r_i, i) \quad (1.8)$$

- N is the total number of queries.
- i represents the index of a query.
- r_i is the rank of the correct match for query i . In other words, it is the position in the ranked list of candidates where the correct match is found for the query.
- $\delta(r_i, i)$ is the Kronecker delta function³, which equals 1 if r_i is equal to i (indicating a correct match) and 0 otherwise.

1.2 Thesis outline

This chapter introduces the technical details and research question, providing the context for the whole study. Chapter 2 presents the results of a literature review on supervised and unsupervised techniques across benchmarks. Chapter 3 discusses research on embeddings of generative diffusion models applied to person re-identification. Chapter 4 covers our investigation of state-of-the-art domain adaptation models. Chapter 5 reports additional

³<https://mathworld.wolfram.com/KroneckerDelta.html>

experiments applying person re-id models to multi-person tracking. Chapter 6 draws the scientific and ethical conclusions.

1.3 Research questions

This thesis is organized around a set of Research Questions (RQs) that guided me during my Ph.D years. Each RQ aims to contribute to the landscape of challenges in person re-identification and is explored throughout different chapters, either partially or fully.

- **RQ1: What are the main open challenges and emerging research trends in supervised and unsupervised person re-identification, according to the latest literature?** This question investigates the state-of-the-art, persistent challenges, and the reasons behind successes and failures of existing approaches in real-world scenarios.
- **RQ2: How can embeddings for person re-identification be extracted to capture only identity-specific features, minimizing influence from transient attributes?** Metric learning often excels at discrimination, but embeddings they produce tend to capture cues like clothing color or illumination rather than intrinsic identity features. This research aims to isolate embeddings representing only inherent, identity-specific characteristics, minimizing transient or irrelevant attributes.
- **RQ3: How robust are current state-of-the-art person re-identification models to domain shifts, and what is the effectiveness of domain adaptation strategies in bridging performance gaps?** This question examines the applicability of models in real-world deployment, analyzing their performance across different datasets, environments, and augmentation methods.
- **RQ4: What is the practical impact of person re-identification models in multi-pedestrian tracking systems, and how can**

their integration be optimized? This question evaluates the role of ReID models in improving tracking accuracy, identifying typical failure modes, and proposing integration strategies for real-world applications.

Table 1.1: Characteristics of some of the person re-id datasets.

Dataset	Year	Identities	Img/Vid	Cameras
CUHK03 [25]	2014	1,360	13,164	6
Market-1501 [3]	2015	1,501	32,668	6
DukeMTMC- reID ^R [4]	2016	1,812	36,441	8
MSMT17 [26]	2017	4,101	126,441	15
Airport [1]	2018	9,651	39,902	6
ENTIRE-ID [5]	2024	13,540	4.45M	37
ENTIRE-ID (Testing Set ¹)	2024	2,741	13,415	37
LUPerson ^U [27]	2021	> 200 <i>k</i>	4.18M	46,260
iLIDS-VID ^V [2]	2014	319	638	2
MARS ^V [33]	2016	1,261	20,715	6
DukeMTMC- VideoReID ^{V,R} [4]	2016	1,404	4,832	8

¹ as of the time of writing, only the test set has been released.^U Unlabeled datasets. ^V Video-based datasets. ^R Retired/withdrawn dataset.

Chapter 2

Literature Review on Person Reidentification

This chapter presents the results of a study conducted in collaboration with Mermec Engineering Srl which involved the literature review process of the most recent techniques applied in Person Re-Identification in the last four years, published in the Machine Vision and Applications journal [14]

The study spanned approximately six months, during which more than 120 publications were reviewed and categorized. The objective was to contribute to the evolving narrative of person re-identification by examining both established supervised techniques and recent advances in unsupervised learning and then be able to compare them.

The next section presents the work of the paper. Unlike the review, where each technique is explained in detail, this chapter follows the same structure but focuses only on the strengths and weaknesses of each family of techniques. It is organized into three sections: supervised, unsupervised techniques and the comparisons of their performances. Finally, the chapter concludes by addressing the research questions and summarizing the key insights from the review.

2.1 Search and selection of papers

The survey was made to compare articles with an available open source implementation, so we started collecting works from Papers with Code¹, which today the platform moved to Hugging Face² - Trending papers. Our specific query targeted “Person Re-Identification”, pointing to the typical benchmarks used to track progress in the field, and comprising “Market1501” [3], “DukeMTMC-reID” [4], and “MSMT17” [26].

We utilized the portal’s capabilities to identify top-performing papers based on metrics such as mAP, Rank-1, Rank-5, and Rank-10.

Then, we manually examined each of these top papers, integrating the source references with previous works and background bibliography in order to obtain a complete and essentially self-contained vision of the domain, and shape a cohesive narrative to unify our findings.

We first categorized the works into supervised and unsupervised methods based on a well-known and widely adopted classification. The articles on supervised methods are clustered into Feature Learning and Metric Learning approaches, according to the emphasis given by the authors to different aspects of their methodology. Works on unsupervised learning are divided into Unsupervised Domain Adaptation, which aims to transfer knowledge from an unlabelled source domain to a labelled target domain, and Fully Unsupervised Learning, which focuses on discriminating between individual images without any prior semantic categories.

2.1.1 Datasets

In this study, we used the already introduced datasets *Market1501* [3], *DukeMTMC* [4] and *MSMT17* [26], as they are the three most widely employed datasets, allowing for a fair comparison across various research contributions. Their information can be found in Table 2.1. All datasets have

¹<https://paperswithcode.com/>

²<https://huggingface.co/papers/trending>

been collected on campus, focusing on local space and short-term settings: the clothes of pedestrians do not change significantly. This assumption has a certain gap with real scenes and would limit the application in real-world scenarios ([13]).

Furthermore, it is important to notice their protocols for computing mAP metric. Conversely to Market1501, in the evaluation of mAP for DukeMTMC and MSMT17, the protocols do not impose any restrictions regarding cameras since there is no overlapping among camera networks.

Dataset	Cams.	(HxWxC)	#Images (IDs)	#Training (IDs)	#Gallery (IDs)	#Query	Ref.
Market1501	6	128x64x3	32,668 (1,501)	12,936 (750)	19,732 (751)	3,368	ICCV 2015
DukeMTMC	8	256x128x3	36,411 (1,812)	16,522 (702)	17,661 (702+408)	2,228	ECCV 2016
MSMT17	15	256x128x3	126,441 (4,101)	32,621 (1,041)	82,161 (3,060)	11,659	CVPR 2018

Table 2.1: Market and Duke dataset information.

2.2 Supervised Techniques

In the context of ReID, the main supervised strategy involves utilizing a labeled training set to train a deep neural network for building a robust representation able to identify individuals. Subsequently, the testing phase involves evaluating the model on previously unseen individuals present in two distinct datasets: the Gallery and Query sets. During the test stage, the network computes the representation of features for each image of the gallery and query sets. This leads to the creation of a distance matrix between every sample of both datasets.

We summarized all the common traits and patterns of the reviewed papers of [14] in a vanilla supervised algorithm showed in figure2.1. This algorithm consists of a neural network that generates embeddings, along with a smaller secondary network used to train the feature encoder using losses strategies such as classification ([16]) and triplet losses ([19]).

According to such representation, two main research approaches play crucial roles, feature learning and metric learning. Both address different aspect

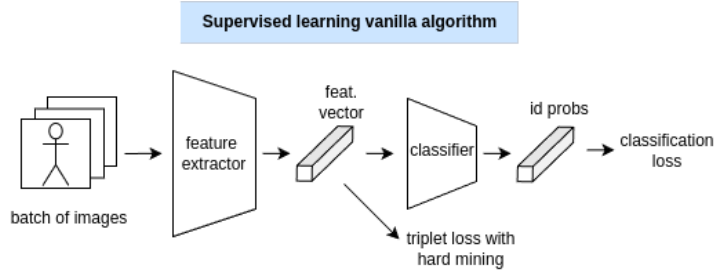


Figure 2.1: Vanilla model for supervised person ReID. During training Triplet loss works by selecting an anchor, a positive (same identity), and a negative (different identity) feature vector from the batch. For classification loss each identity in the training dataset is a class.

of the problem: the first explore the techniques to extract discriminative features from raw images, while the second approach, looking for a similarity metric or distance function that can accurately quantify the similarity between any two feature vectors.

2.2.1 Feature Learning Approach

Feature learning for ReID is a methodology focused on extracting distinctive and discriminative characteristics from raw images, aiming to empower the neural network to effectively capture and represent key information relevant to the correct identification of individuals ([9], [10]).

In this approach, it is common for almost every paper to employ a pre-trained backbone on large-scale datasets (e.g., ResNet50 [56] or Inception V3 [57] trained on ImageNet [53]), leveraging transfer learning to adapt the learned representations to the target Re-ID dataset. Then, specific techniques can be employed on the extracted feature maps to focus on so called *local* or *global features*, discussed hereafter.

Local features refer to fine-grained details and characteristics specific to different body parts, allowing for a more granular representation of individuals. This approach predates deep learning: [58] introduced a deterministic algorithm for vertical and regional symmetry-driven aggregation of local fea-

tures. Today, body part details are learned ([59]–[63]). However, such papers discuss that challenges such as occlusion and background clutter can hinder the accurate extraction of local features and this process often fails to cover all discriminative information, since it almost exclusively focuses on specific parts with fixed semantics.

Conversely, Global features (e.g. [56], [57]) encompass the overall representation of an individual, providing a holistic view. While global features are advantageous for capturing the general appearance of a person, they may face limitations in addressing detailed variations and pose changes, especially in crowded environments.

In the realm of ReID, successful feature-learning strategies ([59], [60]) rely on the combined extraction of both local and global features to leverages the strengths of each type of feature, allowing the model to benefit from the detailed discriminative power of local features while also capturing the general context and appearance through global features.

Typically, the extraction of global and local features occurs within a multi-branch network architecture, where one branch is commonly specialized in extracting global features and the other is dedicated to local features, as in the work by Wang et al. [60].

The Multi-branch networks proved effectiveness for diverse feature extraction. A recent examples that combine both, is the work on MSINet[64] that builds upon OsNet Backbone [65]. It employs multi-scale feature learning by introducing two separate branches at each scale, each with a different receptive field. These branches interact through dedicated modules, and their results are then aggregated to form a comprehensive feature representation. This work is another example of how different receptive fields at different image resolutions can learn different granularity of features within the same architecture, enhancing the discriminative power for person re-identification tasks. This kind of approach in other computer vision contexts led also to the pyramidal approach [66]. Indeed, works such as [60], [64] showed the trade-off between feature richness and computational demands requires care-

ful consideration, impacting real-time efficiency.

Body parts learning The body-part-based approaches are particularly advantageous in scenarios characterized by frequent occlusions or partial visibility, such as crowded environments or surveillance with overlapping fields of view [14]. In such contexts, the ability to focus on visible body regions allows for more robust identification compared to global-only methods, as demonstrated in [67], [68]. However, these methods also highlight that localized feature extraction requires accurate part detection and, especially, suffers from inconsistency across different camera angles due to the variability in appearance and visibility of body parts.

Within the framework of image segmentation, human semantic parsing enhances understanding by finely segmenting and categorizing various body parts like the head, torso, arms, and legs [67], [69]. Atrous Spatial Pyramid Pooling (ASPP), adopted in works such as [69], [70], employs dilated (atrous) convolutions at multiple rates, introducing gaps in the convolutional kernel to extend its receptive field. This technique empowers the network to gather multi-scale contextual information, fostering a more comprehensive understanding of objects at various scales within an input feature map [70]. Furthermore, several methods operate within body parts and multi-branch architectures, combining human parsing and saliency detection to further enhance local feature extraction for person re-identification [67], [68], [71], [72]

Attention-based techniques

Inside the feature-learning, the Attention-based techniques represent a prominent family of techniques in computer vision, designed to emulate the cognitive processes of the human visual system by selectively focusing (zooming in) on the most informative regions or features within an image ([73]). These approaches are also used in other deep learning for tasks such as object detection [34] and image classification [74].

Learned incorporated masks of attention among the feature map of deep

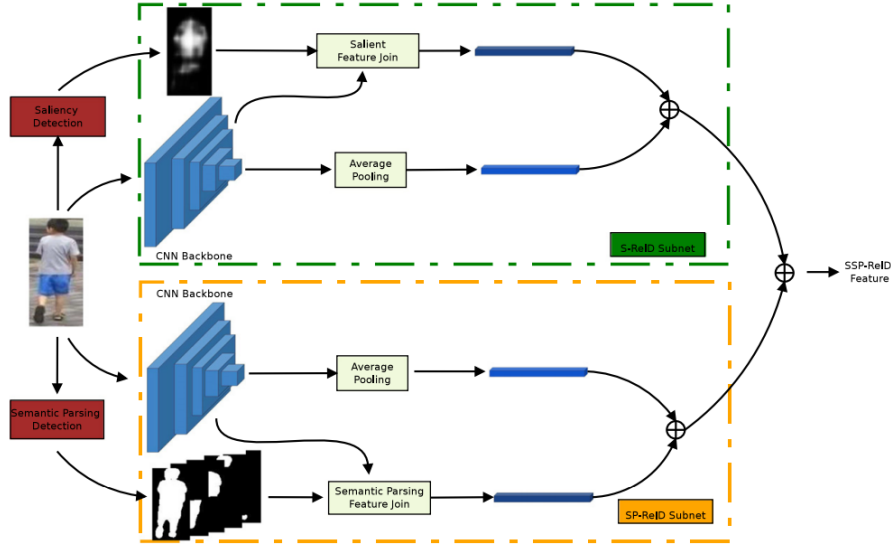


Figure 2.2: SSP-ReID framework. One branch is responsible for human semantic parsing, while the other for saliency detection. Figure taken from [68]

learning models—such as segmentation maps, fully connected layers, or even non-local operators—allow the models to produce attention mechanisms that single out or amplify distinct representations of selected objects directly in the feature map. Non-local operators, in particular, enable the model to capture long-range dependencies by relating distant regions in the image, rather than just local neighborhoods ([75]). All of these mechanisms are used, for instance, in [61], [75]–[77] to address problems such as background clutter, occlusions, scale variations, and feature misalignment. These methods are robust when attention masks correctly isolate foreground features despite such variations, but are less effective when foreground-background separation fails, occlusions are severe, or variations exceed training conditions.

Today, keypoints can also be used as additional information to identify body parts ([78]), avoiding the need to compute a segmentation map. They have the advantage of combining body parts and human pose in a single, manageable cue.

Feature-erasing techniques

Feature-erasing techniques represent a distinct family of methods designed to overcome the limitations of attention-based approaches, which tend to focus on the most salient and easily recognizable regions of the image ([62]). The paper [62] notes that this focus may mask other useful clues, reducing the diversity and richness of learned representations and making it difficult for the network to provide comprehensive and discriminative features for different individuals.

To address these challenges, several feature erasure methods have been proposed ([79], [80]). Common approaches include dual-branch architectures with suppression branches that use binary masks to identify and eliminate the most discriminative features [79], regularization branches to mitigate noise in the erased feature maps [80]. Collectively, these methods force the model to discover complementary information from less obvious or partially occluded areas, ultimately improving robustness and generalization in person re-identification.

Transformer-based techniques

Vision Transformers (ViT) [81] address the feature learning challenges by dividing images into non-overlapping patches. The sequence of patches serves as the input for the ViT, which employs multi-head self-attention within the ViT Block to capture inter-patch representations. The output of a ViT consists of a sequence of tokens, with each input patch assigned a token, alongside a global token, as shown in Figure 2.3. While each patch token encapsulates a representation that integrates both low-level and high-level visual features from its corresponding region in the input image, the global token aggregates information from the entire input image. This approach enables the ViT to capture holistic information about the overall scene, surpassing the limitations of local processing in CNNs and providing a more comprehensive understanding of the input data.

In contrast to the previous ReID techniques where a backbone CNN

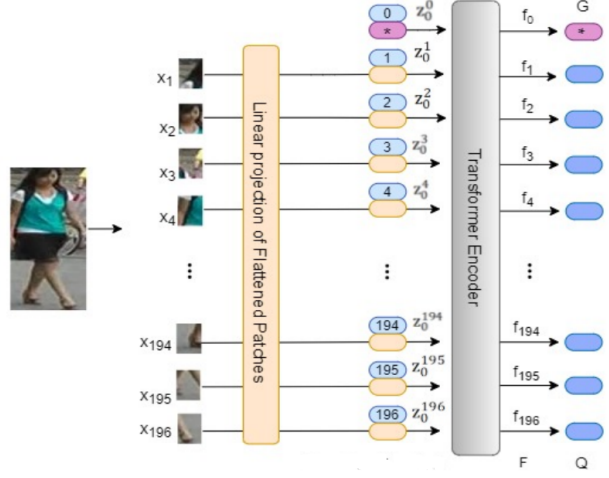


Figure 2.3: In the Vision Transformer (ViT) framework, the image is partitioned into a sequence of non-overlapping patches, forming the input for the ViT model. The model processes these patches, assigning each one a corresponding token. Additionally, a global token, denoted by the letter G and highlighted in violet, represents the overall information from the entire image within the sequence of tokens. Figure taken from [82]

leads to different branches for capturing local/global features, a ViT inherently encapsulates both types of information in the output token sequence. Given ViT's widespread success in various computer vision tasks [83], [84], [85] researchers have increasingly turned to ViT techniques for person re-identification as well, marking a departure from traditional approaches: [86] (Transreid), [82] (LA-Transformer), [87].

In [86] overlapped patches are employed in input, and solely the global embedding is utilized for ID inference.

Recently took place also rigid stripe parts of the input or output of the transformers inspired by PCB [88] work: Sharma, Kapil, and Chapman [82] proposed to arrange ViT encoder's tokens into a 2D grid and processed with row-wise average pooling for classification. Zhang, He, Liu, *et al.* [87], akin to [59], propose to segment the input image into multiple stripes and process them through an encoder with local self-attention and global cross-attention

layers to capture long-range dependencies. Each stripe, represented by a feature vector, aggregates global information from all patch tokens in the pedestrian image.

Cross-Domain adaptation

Cross-domain adaptation refers to the process of adapting a model trained on a source domain to perform well on a different but related target domain. This approach typically assumes some level of supervision in the target domain, which can vary from having a few labeled examples (semi-supervised) to having style or attribute annotations (weakly supervised). We shall discuss fully Unsupervised Domain Adaptation in Section 2.3.1.

Cross-domain adaptation methods often involve:

- **Transfer Learning:** Leveraging knowledge from the source domain to improve performance in the target domain. This might involve fine-tuning a pre-trained model on a small set of labeled data from the target domain [89], [90].
- **Domain Mapping:** Transforming data from the source domain to appear as if it came from the target domain (or vice versa), thereby reducing the domain gap. Generative models, like GANs, are commonly used for this purpose [91]–[93].
- **Feature Alignment:** Modifying the feature space to minimize the discrepancy between source and target domain distributions in the feature space. This process involves adjusting the features extracted from both domains so they become more similar, making it easier for a machine learning model to perform accurately on the target domain without requiring extensive labeled data from that domain [94], [95].

Combining cross-domain adaptation with data augmentation offers a synergistic approach to handling domain shifts. Data augmentation can increase the diversity and volume of training data, simulating conditions that may

not have been covered by the original dataset. This enriched dataset can then serve as a better foundation for cross-domain adaptation techniques, which fine-tune the model to align more closely with the target domain’s characteristics.

For example, generative models like GANs or Diffusion Models [96], [97] can be used not only for feature alignment but also for generating synthetic images that bridge the gap between the source and target domains. These images can then be used as part of the training set, effectively augmenting the data with examples that are directly relevant to the target domain’s distribution.

A different perspective on the use of generative techniques for Person-ReID has been recently advocated in [98], and also we measured supervised domain adaptation in [99]. We shall discuss them in the next chapters (Chapters 3 and 4) of this document.

2.2.2 Metric Learning Approach

The Metric learning approach in the realm of Person Re-Identification is the approach of employing specific loss functions to guide the feature extractor into learning to cluster identities correctly.

Among the various approaches, metric learning enables models to adapt to open-world scenario assumptions, where the goal is to re-identify unseen identities during testing. Without this capability, it would not be possible to effectively handle new or unseen classes in real-world applications (as analyzed in [16], [19]).

We summarize this section based on the review [14] into four parts. The first two sections, 2.2.2 and 2.2.2, present the two most commonly used families of loss functions. Section 2.2.2 discusses recent general advances, while section 2.2.2 covers the application of contrastive learning using CLIP-based models.

Classification-based losses

The Cross-entropy loss (*CE-loss*), has shown a general effectiveness into separate classes [15], [22]. In the realm of Person ReID, every person in the training set is a class and the network learns to separate such classes. During inference, the classification layer (last) is removed, and the output of the penultimate layer is used as the embedding for re-identification [15], [100].

The *CE-loss*, in literature also *Identity loss*, encourages the model to learn semantically rich embeddings that effectively separate different identities. Additionally, applying label smoothing during training is a well-established practice to enhance generalization ([15], [101], [102]).

$$L(p, q) = - \sum_{i=1}^C p_i \cdot \log(q_i) \quad (2.1)$$

where:

- p_i is the true probability of the input belonging to class i .
- q_i is the predicted probability of the input belonging to class i .
- C is the number of identities .

This loss doesn't has limitation in open-world scenarios where unseen identities appear [15]. To address this problem, it is common to combine classification loss with a *Verification strategy*: the binary classification problem (Person verification) given a pair of images. This latter uses binary cross-entropy loss and together can enhance model's performance by leveraging their respective advantages: while the classification loss creates separation among distinct classes, the verification loss minimizes the distance within the same class; did so [103], [104].

$$L_{\text{BCE}} = - \frac{1}{N} \sum_{i=1}^N [y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] \quad (2.2)$$

where:

- N is the number of sample pairs.
- y_i is the binary label indicating whether the pair is of the same person (1 if yes, 0 if no).
- p_i represents the predicted probability of the pair being of the same person.

Contrastive-based losses

Contrastive learning is a fundamental component of metric learning applied to person re-identification. The core idea is to minimize the embedding distance between images of the same identity (intra-class distance) while maximizing the distance between different identities (inter-class distance). Among contrastive losses, the triplet loss framework is predominant, often used alongside identity loss to improve feature clustering [16], [105]. Given an image embedding $z_i = f_\theta(x_i)$ produced by a network f_θ , a triplet consists of an anchor z_a , a positive z_p (same identity), and a negative z_n (different identity). The standard triplet loss is formulated as:

$$L_R(z_a, z_p, z_n) = \max(0, m + d(z_q, z_p) - d(z_q, z_n)) \quad (2.3)$$

Since the input images are normalized to fall within the $[0, 1]$ range, the output resides on a hypersphere, $d(\cdot)$ is the Euclidean or Cosine distance and m is a margin enforcing separation. Hermans, Beyer, and Leibe [16] advanced this by proposing hard-mining of positives and negatives within batches to create more informative triplets:

$$L_R(z_a, z_p, z_n) = \max(0, m + \max_{p=1\dots k} (d(z_q, z_p)) - \min_{n=1\dots k} d(z_q, z_n)) \quad (2.4)$$

Further enhancements include weighting negative samples using focal loss [106] and incorporating centroid-based regularizers to improve cluster com-

pactness [105], [107]–[109]. Alternative generalizations like quadruplet [110] and N-tuple losses [111] have also been proposed.

Effective triplet sampling is crucial due to the combinatorial explosion of possible triplets, with strategies developed to select hard or moderately informative triplets [16], [112], [113], ensuring efficient and stable training.

Beyond the usual losses

Recent parallel advances in metric learning losses such as the *Angular Losses* family including ArcFace loss [100] and NT_Xent [114], have also benefited person re-identification.

Some works have pushed beyond traditional losses by integrating them with data augmentation or transformation techniques [115], [116], incorporating temperature-scaled losses [117] like InfoNCE [118], or applying triplet angular losses [18]. These methods, often combined with network modifications [115], [116], address typical learning challenges by encouraging more discriminative and robust features. Despite these innovations, notable improvements on standard benchmarks remain few.

ProNet++ [119] presents improvements in triplet losses and inference techniques: it projects features onto the classifier weights, aligning them with the supervised training space and improving the semantic consistency of retrieval descriptors, especially under pose variations and occlusions. It achieves state-of-the-art results on the Market-1501 and MSMT17 datasets.

CLIP-based Person ReId

A recent and promising approach involves leveraging advanced image-to-text embedding techniques to group images based on their visual similarities. A notable tool in this domain is OpenAI’s Contrastive Language–Image Pre-training (CLIP) [120]. CLIP can be adapted to both supervised and unsupervised settings, depending on its application.

In a supervised context, CLIP is fine-tuned on a labeled dataset to master the specific task of matching different images of the same individual, making

use of its pre-trained capabilities on a vast corpus of image-text pairs (i.e. [121]–[123]). It can be used in normal supervised settings with a two-stage training, the first train the text-decoder to describe an identity and the second frozen this latter and train the image-decoder. Another uses case was in the context of video-based person reid [124]; and in unsupervised framework to generate pseudo labels through clustering of CLIP embeddings and fine tune with them the labels (i.e. in [125]).

Clothing-Change Person ReID (CC-ReID)

As already stated, the Clothing-Change Re-ID task recognizes individuals across different outfits, relying on invariant biometric and appearance cues [20]; This study primarily uses standard short-term ReID benchmarks where each individual typically appears in a single outfit; dedicated datasets such as PRCC [126] and LTCC [12] are specifically designed to assess robustness to clothing changes. However, techniques from the feature and metric learning families that has been presented so far remain valid and complementary for the CC-ReID task. Examples include: dual-branch architectures separating identity and clothing features [127]; Vision Transformers combined with body-part semantic segmentation [128]; integration of face and body features [129]; gait-based modules with contrastive loss for individuals wearing different outfits [130]; and GAN-based generation of synthetic images with varying clothing and poses [131].

2.3 Unsupervised Techniques

The unsupervised techniques for person re-identification can be broadly categorised into two main tasks: unsupervised domain adaptation (UDA) and unsupervised image encoder learning also known as fully unsupervised learning. UDA typically requires labeled data in the source domain but does not require labeled data in the target domain; fully unsupervised learning aims to learn useful patterns or representations directly from unlabeled data

without any domain adaptation. We categorize such approaches below as done in [14].

2.3.1 Unsupervised Domain Adaptation methods

Unsupervised Domain Adaptation (UDA) consists of pre-training a model on labeled source-domain data and then adapting it to an unlabeled target domain to transfer knowledge and improve performance.

After pre-training, the model backbone extracts features from target-domain samples. Clustering algorithms such as K-means [132] or DBSCAN [133] are applied to these features to assign pseudo-labels, which act as surrogate supervision. These pseudo-labels, though not ground truth, enable the use of supervised objectives like metric or contrastive loss to further refine the model on the target domain [134].

After pre-training, the backbone extracts features from unlabeled target samples. A clustering algorithm (e.g., K-means [132], DBSCAN [133]) generates pseudo-labels for these samples, introducing supervision. These pseudo-labels, not ground truth, guide further training by enabling metric or contrastive loss computation [134].

The UDA clustering methodology alternates between generating pseudo-classes through clustering target-domain instances and training the network with them, allowing adaptation by capturing inter-sample relations in the target domain [134]. For instance, they can be employed to calculate metric or contrastive loss, which necessitate labelled data to supervise the learning process.

Nevertheless, in this approach the main challenge lies in pseudo-label quality, as noise arises from:

- Domain gaps—differences in identities, viewpoints, lighting, and backgrounds [135].
- Clustering characteristics—coarse hard labels grouping diverse samples

under one identity [136], increasing misclassifications risks and so, reducing robustness.

Mitigating noisy pseudo-labels is crucial in UDA for ReID [137]. A common solution that allows for more flexible uncertainty modeling and robust adaptation, used also in [137], is to use soft pseudo-labels—probability distributions over classes—rather than hard assignments. These can be computed with a fully connected layer, refined using one-hot encoding and cross-entropy to align soft and hard labels.

Recent strategies address noisy pseudo-labels in UDA ReID by improving label quality. The following sections detail these techniques.

Mean Teacher-Student

The **mean-teacher framework** [138] is a semi-supervised learning approach involving two networks — a *student* and a *teacher*. Both start with identical architectures and receive the same input with different noise. The student updates via gradient descent, while the teacher updates its weights through an *exponential moving average (EMA)* of the student’s weights, which stabilizes training and produces more reliable predictions. The EMA update strategy improves accuracy compared to the teacher directly sharing weights with the student [139]. When training end, both models can predict, though the teacher’s output is more reliable, as it averages the optimized student model with its past parameters.

In the realm of UDA for ReID, the mean teacher-student framework stands out as a widely embraced backbone feature extractor [140], [141], [142], [143], [144] due to its effectiveness in leveraging the teacher network’s reliability to improve the generation of pseudo-labels, facilitating more robust feature extraction and so, effectiveness in cross-domain Re-ID.

Further this framework is mainly applied to:

- **Generate more reliable pseudo-labels through teacher network predictions:** Mutual Mean Teaching (MMT) [140] employs

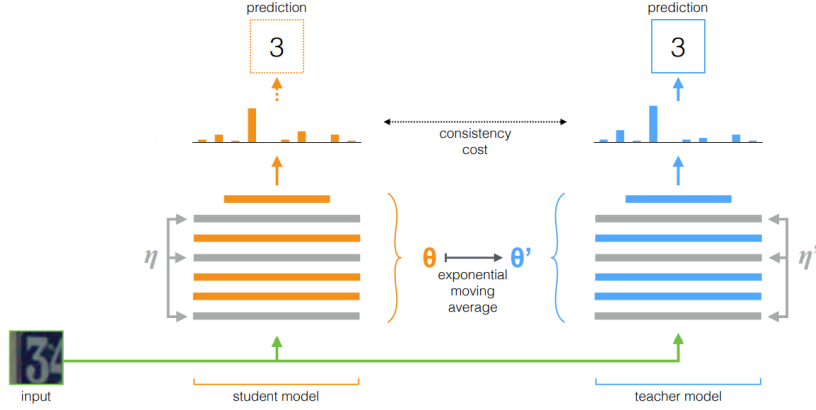


Figure 2.4: In the mean teacher-student framework, both models output the prediction for the same input (corrupted with different noise η and η'). Then the consistency loss is calculated, to align the two predictions. Finally, the student model is updated via gradient descent, while the teacher model via exponential moving average (EMA). Figure taken from [138]

two mean teacher-student networks that exchange soft labels to refine pseudo-labels via a modified triplet loss and soft cross-entropy consistency.

- **Improve feature extraction in cross-domain scenarios and reduce noise in label assignment and clustering processes:** JAC-Net [141] integrates two mean teacher-student networks with a joint learning network to leverage combined knowledge, enhance discriminability, and mitigate noise expansion during training.

These methods present some limitations, primarily due to the increased computational cost associated with maintaining dual networks. They also require careful tuning of EMA and loss parameters to ensure stability and effectiveness. Furthermore, there is a risk of confirmation bias if pseudo-labels are not updated properly, and the overall complexity grows with advanced extensions such as the work presented of MMT[140] and JAC-Net[141].

Memory-based contrastive techniques

Memory-based contrastive techniques utilize memory banks to retain feature vectors computed during training, preserving the entirety of training data feature vectors. This enables the application of a clustering algorithm on the complete dataset, distinguishing it from non-memory approaches that solely cluster samples within each batch. ([145]). Mining of hard positive and negative examples from the whole dataset can be done, improving contrastive power and pseudo-label reliability. Without it, loss computation is limited to batch samples, reducing diversity and possibly adding noise to pseudo-labels, which harms training. Variations in memory-based methods mainly concern how memory is managed, affecting stored feature vectors and the samples used for contrastive loss computation. ([145]).

In a standard framework, a clustering algorithm is applied to all image features in the memory bank to generate pseudo-labels ([145]). The model is then trained with a contrastive loss, typically the softmax form of the InfoNCE loss ([118]). For a given set of N samples, the softmax InfoNCE loss of a sample x_i belonging to the set is computed as:

$$L_{\text{InfoNCE}} = -\log \left(\frac{\exp(\text{sim}(x_i, x_j)/\tau)}{\sum_{k=1}^N \exp(\text{sim}(x_i, x_k)/\tau)} \right) \quad (2.5)$$

where x_j is a positive sample from the set, $\text{sim}(x_i, x_k)$ is a similarity score (e.g., cosine similarity) between the pair (x_i, x_j) , and τ is a temperature parameter. The choice of which feature vectors are stored in memory directly influences which samples are considered in the contrastive loss ([145]). Some methods use two separate memory banks or additionally store centroids, batch samples, and outliers, enabling auxiliary contrastive losses and specialized sampling strategies ([146]; [145]). Typically, a DBSCAN update is performed on the memory at each training epoch to update pseudo-labels and centroids ([145]–[147]).

Feature Learning

Feature learning in UDA for Person ReID focuses on extracting discriminative global and local features from unlabeled images. The typical pipeline such as the one in Fu, Wei, Wang, *et al.* [147], involves feature extraction (global + horizontal-stripe local) with ResNet-based backbones, pseudo-label generation via clustering and compute triplet loss independently for each feature type. Ding and Zhou [143] address the problem of "obscure learning" by fusing global and local features in the LF² framework. Memory-based contrastive methods, e.g., MGMDG [144], use dedicated memory banks for features at multiple granularities to enhance feature robustness, often combined with mean teacher-student architectures [140].

Overall, common techniques in UDA feature learning include clustering-based pseudo-labels, feature fusion to reduce ambiguity, memory-based contrastive learning, and teacher-student frameworks for stability.

Self-paced learning

Self-paced learning (SPL) is a learning paradigm that adjusts the learning process by gradually introducing training samples based on their difficulty. This approach is inspired by curriculum learning [148], which mimics human learning by first mastering easy samples and then gradually adding harder ones. In unsupervised ReID, *SPL* helps guide learning without labeled data, reducing noisy pseudo-label effects and improving robustness ([137], [149]–[151]).

The approach suits person re-identification, where individual appearances vary significantly across domains and scenarios [149], [150]. Simple *SPL* techniques rank samples by difficulty based on predefined conditions to guide training [148].

This rationale is also used:

- to gradually shift from the source domain (simple labeled samples) to target domain (hard, noised pseudo-labels) [149].

- to filter out hard/noisy samples using clustering (DBSCAN) and prioritize training on inliers before introducing harder outliers [150].
- to compute a reliability score and split clusters into easy, moderate, and hard subsets, training progressively from easier to harder samples [151].
- to dynamically adapt clustering parameters and filter noisy pseudo-labels across cameras, improving robustness in cross-camera settings [137].

Based on the observations in [149]–[151], it can be further generalized that the *SPL* approach is highly sensitive to the choice of difficulty criteria; if these are not properly balanced, the model may underfit harder samples, and large domain shifts can make the definition of difficulty unreliable.

2.3.2 Fully unsupervised methods

Unsupervised methods work with unlabeled datasets, conversely to the supervised methods, learning to discriminate individual images without any notion of semantic categories. Thus, given a training set: $X = \{x_1, \dots, x_n\}$, the objective is to develop a model $F(\theta, x_i)$ where θ represents the parameters of the model and the function F maps input images x_i to their corresponding feature representations. A good model should map images of the same person closer to each other while pushing images of different identities further apart. The inference phase typically follows the same approach as supervised techniques: ranking the distances between a query and all gallery images and the test is performed over a supervised dataset.

The common baseline algorithm for training with unsupervised reidentification approach combines self-supervised clustering and contrastive learning. Features are extracted, clustered by similarity, and centroids computed as the mean of its embedded vectors. During the training the new image features are then compared to centroids as positive or negative sample. This is schematically described in Figure 2.5. The model maintains

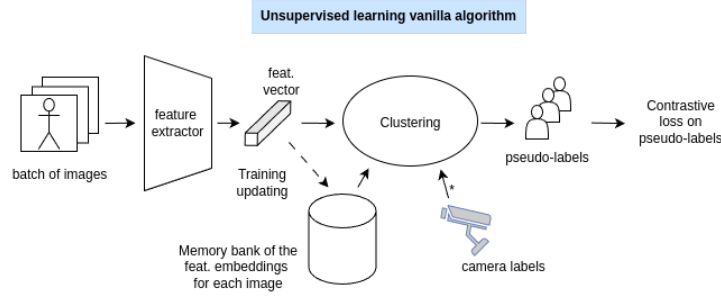


Figure 2.5: Vanilla model for unsupervised person ReID. The extracted features are stored in a dictionary for each image. These embeddings, combined with camera labels and a clustering algorithm, generate pseudo-clusters, which are then used to optimize contrastive losses. Text and image taken from [14]

a memory bank storing feature embeddings for each image in the dataset. This helps in comparing current features with past features, enabling better clustering and pseudo-label generation for unsupervised learning. Camera labels and additional information, such as pose estimation, can be optionally included to enhance clustering by accounting for different perspectives. ([14])

Feature learning

Choosing the appropriate backbone architecture for extracting feature vector at different granularities remain a recurrent problem also in the domain of fully unsupervised person re-identification learning.

The most used backbone here are RN50 [152], [153], and ViT [154], [155] based as well the multi-branch form of the networks and what derives from that. Further, is important to notice also the usage of data augmentation [153] can differ for each branch, and stripes strategies of the output of the ViT and the usage of camera information inside the tokens [154]. Pre-training ViT models in the unsupervised dataset LuPerson dataset [152] can improve performance on supervised benchmarks like Market and Duke. This was shown in [155].

Anyway the main limitation is the significant differences between source and target distributions that can lead ViT suffering of catastrophic forgetting

problems leading to poor results. To mitigate this, the authors of [155] introduced a Catastrophic Degradation Score to assess distribution similarity and developed a filtering method that reduces the source dataset by 50% of samples while maintaining benchmark performance.

Effective Clustering & Sampling

Designing effective clustering and sampling strategies without labels is challenging. Clusters may merge distinct identities or split the same identity, and poor sampling can waste learning on irrelevant or redundant data ([156]–[158]).

A common approach proposed in the work referenced in this section, is to initialize a dictionary before training using clustering results—pseudo-labels, centroid feature vectors, outliers, camera data, and other relevant information—which is updated after each training epoch. These clustering-assigned labels are known as pseudo-labels.

The effectiveness of clustering during training depends on several key factors:

- **Memory Bank Content:** The use of global, local, and combined features [156], as well as camera-based proxies [159], [160], enriches representations by incorporating diverse information. However, this approach can increase complexity and introduce noise or bias.
- **Sampling Strategy:** Approaches such as hard sample mining [156], [160], stochastic sampling with updated centroids [158], time-based proxy selection [159], and online/offline inter-camera contrastive losses [161] focus training on the most informative samples. Nonetheless, they may increase computational cost and risk introducing sampling imbalance.
- **Other Innovations:** Strategies including two-stage intra-camera and inter-camera training [157], as well as support sampling with linear interpolation and label-preserving loss [162], improve robustness and

mitigate pseudo-label noise. These innovations, however, add complexity and require careful design considerations.

Learning from noised pseudo-labels

As discussed in [142], [163]–[166], in self-supervised clustering and contrastive learning, noise arises when incorrect labels or representations are acquired, leading to errors in clustering or similarity calculations and impacting the overall model performance. To mitigate this effect, relying solely on camera information is insufficient. Recent works have employed a combination of techniques based on student-teacher and momentum updating strategies [163], [164].

In the context of unsupervised person re-identification, the momentum strategy, as proposed in [163], involves updating the feature extraction memory according to the "speed" of updates, smoothing the feature extraction process, enabling gradual adaptation, and potentially enhancing convergence during training.

Strategies to learn from noised pseudo-label are:

- **Stabilize learning via momentum and teacher–student models** [163], [164]: Momentum-based optimization and separate functional modules for feature extraction and label noise mitigation enhance training stability and robustness.
- **Align teacher and student representations** [142]: Techniques that align projections from both student and teacher networks toward pseudo-class centers using sharpening techniques improve convergence and reduce noisy labels.
- **Incorporate inter-camera samples using Camera-Aware Jac-card Distance** [165]: Methods that address noise by incorporating inter-camera samples and reweighting them via Camera-Aware Local Query Expansion improve label reliability.

- **Correct pseudo-labels based on confidence estimation** [166]: Inhibition frameworks that actively correct erroneous pseudo labels based on confidence scores improve robustness.

Note: Each modification proposed adds complexity, computational cost, and may require careful parameter tuning, potentially affecting scalability and generalization.

Instances Relationship Understanding

This section presents methods for distinguishing intra-class variations (pose, lighting, viewpoint) from inter-class similarities without supervision. The following works tackle this problem using memory updates, graph structures, attention mechanisms, and multiple pseudo-labels:

- **Stochastic Real-time Memory Updating** [167]: randomly replaces a memory dictionary entry with the current minibatch feature. Can cause instability and requires careful update tuning.
- **Graph-based Intra-class Modeling (DiDal)** [168]: uses a graph to model local/global relationships, handling occlusions and pose variations. Graph construction is computationally costly.
- **Affinity-based Attention** [169]: builds embeddings with a similarity matrix $R_{i,j}$ to capture cluster correlations. Extra computation and attention weight calibration is required.
- **Dynamic Attention Allocation (SimAM)** [170], [171]: assigns 3D attention weights based on image content. Increases computational cost and parameter tuning.
- **Multiple Pseudo Labels** [172]: generates multiple pseudo-labels per sample, refining clusters and updating memory adaptively. More pseudo-labels can increase noise, requiring robust filtering.

- **Teacher-Student Alignment** [172]: aligns a student network to a teacher to resist noisy labels. Adds complexity and depends on teacher quality.

Unsupervised Lifelong Learning with Anti-forgetting Adaptation

Methods for unsupervised lifelong learning in Re-ID models, focusing on retaining previously learned representations while adapting to new data. To tackle this task contrastive learning, pseudo-labels, meta-learning, and memory-based mechanisms can be used: in the DJAA [173] framework an Adaptation module learns representation of new instances via contrastive losses guided by clustering-based pseudo-labels while a Rehearsal Module enforces similarity consistency with a frozen previous model using a memory buffer of past instances and cluster prototypes to face the anti-forgetting problem. The framework employs a dual-level approach (image-to-image and image-to-prototype) on the loss of both its modules. LSTKC [174] proposes an adaptive and unsupervised anti-forgetting mechanism that calculates a single global coefficient to estimate the divergence between the old and the updated model. This calculation is based on the element-wise difference between the normalized similarity matrices (Relation Matrices) of both models. This coefficient dynamically regulates the parameter fusion between the two models, assigning a higher weight to the old model when the divergence is high, thus preserving the acquired knowledge. The CLUDA framework [175] coordinates lifelong learning via a meta-learning strategy: the Meta-Train phase adapts to unlabeled data using a self-supervised loss based on DBSCAN-generated pseudo-labels, while the Meta-Test phase ensures anti-forgetting by aligning gradients to prevent degradation on historical data. Concurrently, the Relational Consistency Learning (RCL) module preserves ranking structures by distilling inter-instance affinity relations from a momentum-updated historical teacher rather than just logits. To support this, the memory buffer employs ID-wise reservoir sampling to maximize the diversity of stored identities.

Human-Centric Perception

Human-Centric Perception [176], [177] focuses on interpreting human-related visual data, unifying tasks like person re-identification, pose estimation, and human parsing. This holistic approach is crucial for addressing complex open challenges where relying on appearance alone is insufficient such as Clothes-Changing Re-Identification (CC-ReID) or unsupervised life-long learning with anti-forgetting adaptation. The main challenges include handling diverse appearances, varying poses and viewpoints, and the interdependence of tasks. For example, SOLIDER [176] highlights the need to balance semantic and appearance information via a semantic controller to optimize performance across tasks. UniHCP [177] addresses multiple tasks using a plain vision transformer, where the same features are processed and then decoded into different outputs for each task.

2.4 Performance Analysis and Comparisons

In this section, we first summarize the reported performances of supervised and unsupervised methods separately, and then provide a comprehensive comparison between the two.

2.4.1 Supervised ReID performance

First, we are going to present and discuss the results of supervised techniques. Table 2.2 illustrates the best 14 publications on the Market1501 dataset presented in this survey, reporting also the performances on the Duke dataset. Among these methods, those also evaluated on MSMT17 are reported in Table 2.3. Notably, papers employing the combination of triplet and classification loss demonstrate stronger performance across both datasets, surpassing other methods. Globally, all these works exhibit mAP well above 90%, with almost all hovering around 90%. Compared to Market1501, the available performances on DukeMTMC are slightly lower but

still above 90%. Specifically, [109] stands out, achieving the top-tier results on both Market1501 and DukeMTMC, leaving minimal room for further improvement.

	Market1501		DukeMTMC		
Paper	mAP	Rank-1	mAP	Rank-1	Loss
[176] Solider	93.9	96.9	-	-	CL
[82] LA-Transformer	94.5	98.3	-	-	CL
[117] VRN	95.2	96.3	-	-	CO
[64] MSINet	89.6	95.3	-	-	CLTR
[86] TransReID	89.5	95.2	82.1	91.1	CLTR
[105] Wieczorek	98.3	98.0	92.7	94.5	CLTR
[178] stReid	95.5	98.0	92.7	94.5	CLTR
[179] LightMBN	91.5	96.3	-	-	CLTR
[115] FlipReID	94.7	95.8	90.7	93.0	CLTR
[116] LDS	94.9	96.2	92.9	91.0	CLTR
[109] MctI	98.6	98.4	97.3	96.8	CLTR
[123] Clip-ReID	90.5	95.4	83.1	90.8	CLTR
[119] ProNet++	95.3	96.4	-	-	CLTR
[165] CA-Jaccard	94.5	96.2	-	-	CLTR

Table 2.2: Best-supervised person re-identification publications reporting at least Market1501 are included; '-' indicates that Duke performance is not reported in the original work. On the last column, CL stands for classification loss, 'TR' for triplet, 'CO' for contrastive, and 'CLTR' for the combination of classification and triplet loss. The best work is highlighted in bold.

2.4.2 Unsupervised ReID Performance

In the realm of unsupervised person re-identification, it is noteworthy to initially compare the two presented UDA techniques before delving into a comparison with fully unsupervised methods. However, the evaluation

	Market1501		MSMT17		
Paper	mAP	Rank-1	mAP	Rank-1	Loss
[176] Solider	93.9	96.9	86.5	91.7	CL
[86] TransReID	89.5	95.2	69.4	86.2	CLTR
[115] FlipReID	94.7	95.8	81.3	87.5	CLTR
[116] LDS	94.9	96.2	79.09	88.35	CLTR
[123] Clip-ReID	90.5	95.4	86.7	91.2	CLTR
[165] CA-Jaccard	94.5	96.2	74.1	86.2	CLTR
[119] ProNet++	95.3	96.4	80.0	88.2	CLTR

Table 2.3: Best-supervised person re-identification publications reporting results on both Market1501 and MSMT17 are included; In the last column, CL stands for classification loss and CLTR for the combination of classification and triplet loss. The best-performing work is highlighted in bold.

of Domain Translation methods introduces a challenge, as it involves different datasets as source and target domains, complicating a fair comparison. Moreover, their performances consistently fall short when compared to UDA Clustering-based methods. Given these considerations, we opt for a direct comparison between the top five publications in UDA clustering-based methods and the top five publications in fully supervised methods on the Market1501 dataset, reporting also performances on DukeMTMC. This evaluation is based on their performances on Market1501 and DukeMTMC, as detailed in Table 2.4. Notably, for UDA clustering-based works, the 'Market1501' column reflects the dataset as the target and DukeMTMC as the source, while the 'DukeMTMC' column represents the reverse scenario (DukeMTMC as the target and Market1501 as the source).

This table highlights the robustness of UDA clustering techniques for each dataset. In fact, on the Market1501 dataset, the mAP performances are consistently around 85%, while on the DukeMTMC dataset, they range between 70% and 75%. However, the gap between the two datasets is quite evident, reaching a difference of at least 10% for each work. Regarding fully unsu-

	Market1501		DukeMTMC		
Paper	mAP	Rank-1	mAP	Rank-1	Technique
[137] ZhaoRN50	84.1	92.9	74.8	86.1	UDAC
[146] DPCFG	85.4	94.2	73.7	85.7	UDAC
[142] HDCRL	83.9	93.6	70.9	83.5	UDAC
[143] LF ²	83.2	92.8	73.5	83.7	UDAC
[151] DUCL+PAFR	84.0	93.9	72.7	84.9	UDAC
[165] CA-Jaccard	86.1	94.4	-	-	FUNS
[157] Repel+Absorb	87.0	94.5	80.3	89.9	FUNS
[154] TMGF _{LP}	89.5	95.5	76.8	86.7	FUNS
[162] ISE	84.7	94.0	-	-	FUNS
[152] LupRN50 _{LP} (SpCL)	76.2	90.2	67.1	81.6	FUNS
[155] CFS _{LP}	89.6	95.3	-	-	FUNS

Table 2.4: Performance of top unsupervised person re-identification publications both for UDA and fully unsupervised techniques. If the performance of a dataset is not reported in the original work, we use the symbol '-'. On the technique column, 'UDAC' stands for UDA clustering, and 'FUNS' for fully unsupervised. The best work is highlighted in bold. All 'UDAC' methods use *Market1501* and *DukeMTMC* alternately as source and target datasets. *LP* means that a model is pretrained with LUPerson dataset [27].

pervised techniques, here too, the performances consistently exceed 85% on the Market1501 dataset, while the lack of data on the DukeMTMC dataset prevents a direct comparison. Finally, when comparing fully unsupervised techniques with UDA clustering techniques, we notice that the former consistently exhibit higher performances, with [152] showing the best results thanks to a mAP of 92%.

2.4.3 Method Comparisons

Finally, we compare the top 5 supervised and unsupervised works. The outcomes are detailed in Table 2.5. As expected, supervised techniques demonstrate superior performance on both datasets compared to their unsupervised counterparts, with [109] emerging as the top-performing work once again. Supervised methods consistently achieve a mAP on the Market1501 dataset of approximately 95%, with [105], [109] surpassing 98% on Market1501. It is noteworthy that [109] also improves performance on DukeMTMC, reaching a mAP of 97%. These significant results highlight the near saturation of performance for supervised techniques on both datasets, leaving very little room for improvement.

Unsupervised approaches, on the other hand, range from 86% to 90% on the Market1501 dataset, except for [152] and [154], which reach 92%. The observed gap between the top supervised and top unsupervised methods is about 6% on the Market1501 dataset. This difference is not far from being considered acceptable, especially when taking into account the absence of labelled data in training and achieving mAPs around 92%. Furthermore, these unsupervised works, unlike supervised ones, still have room for improvement. The scenario differs when we compare results on DukeMTMC, where the gap with supervised techniques is still significant (approximately 13%).

2.4.4 Conclusions

In the presented work, the state-of-the-art methods for both supervised and unsupervised person re-identification has been compiled and analyzed.

For the supervised approaches, we categorized the methods based on the loss functions they utilize, and our comparison highlights that the combination of classification and triplet loss consistently yields the best results. For the unsupervised approaches, we collected the top-performing methods from the past three years and classified them into two categories: Unsupervised Domain Adaptation and Fully Unsupervised methods. Our comparison

	Market1501		DukeMTMC		
Paper	mAP	Rank-1	mAP	Rank-1	Technique
[105] Wieczorek	98.3	98.0	96.1	95.6	SL
[117] VRN	95.2	96.3	-	-	SL
[178] stReid	95.5	98.0	92.7	94.5	SL
[116] LDS	94.9	96.2	92.9	91.0	SL
[109] Mctl	98.6	98.4	97.3	96.8	SL
[157] Repel+Absorb	87.0	94.5	80.3	89.9	USL
[154] TMGF	89.5	95.5	76.8	86.7	USL
[162] ISE	84.7	94.0	-	-	USL
[152] LupRN50(SpCL)	76.2	90.2	67.1	81.6	USL
[155] CFS	89.6	95.3	-	-	USL
[137] ZhaoRN50	84.1	92.9	74.8	86.1	USL
[146] DPCFG	85.4	94.2	73.7	85.7	USL

Table 2.5: Performance of top-5 supervised and unsupervised person re-identification publications on both datasets. If the performance of a dataset is not reported in the original work, we use the symbol '-'. On the technique column, 'SL' stands for supervised learning, and 'USL' for unsupervised learning. The best works for each paradigm are highlighted in bold.

indicates the superiority of fully unsupervised approaches in this domain.

We then conducted a comparative analysis of the performance of both learning paradigms on two widely used person re-identification datasets: Market1501 and DukeMTMC. The state-of-the-art supervised methods demonstrate outstanding performance on both datasets, achieving around 98%, thereby indicating that further improvements in this area are likely to be minimal. In contrast, state-of-the-art unsupervised methods show varying levels of performance depending on the dataset. On the Market1501 dataset, the performance gap between the two paradigms is approximately 6%, which is relatively small, especially considering that unsupervised methods do not rely on labeled data and still achieve around 92% performance. However, on

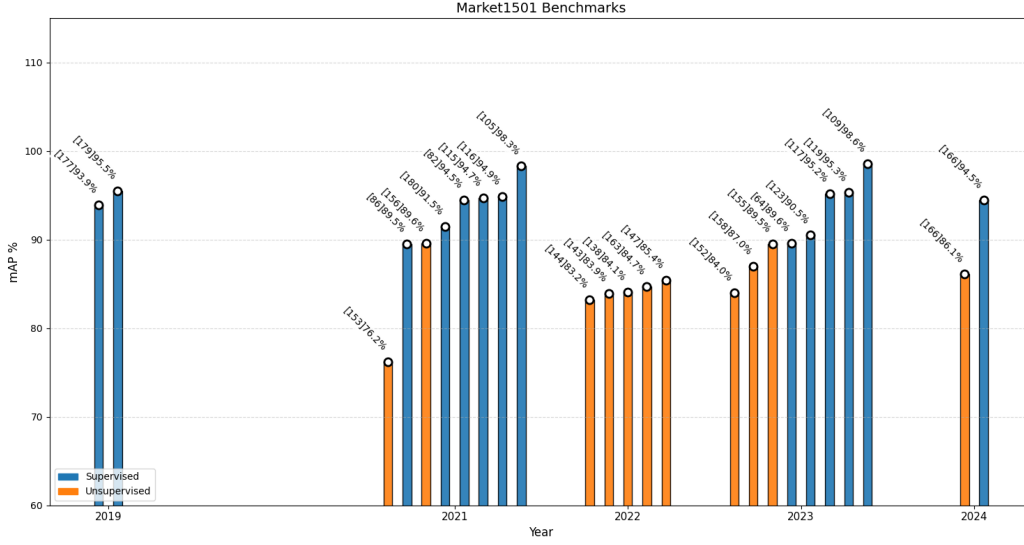


Figure 2.6: This figure illustrates the mean Average Precision (mAP) of the Market1501 dataset for both supervised and unsupervised techniques among years. The results highlight the performance comparison between these two categories of methods. Note that DukeMTMC and MSMT17 benchmarks are not included, as many models were not evaluated on this dataset.

the DukeMTMC dataset, the performance gap is more pronounced, amounting to 13%.

In summary, our analysis demonstrates the near saturation of supervised performance on both the Market1501 [3] and DukeMTMC [4] datasets. Additionally, we highlight the commendable performance of unsupervised techniques on the Market1501 dataset, where the gap with supervised methods is relatively narrow. Conversely, there remains a substantial performance gap between supervised and unsupervised methods on the DukeMTMC dataset.

2.5 Conclusions on Research Questions

This section presents an analysis of the research question, supported by the data of the discussed paper [14]. For the sake of completeness, the question are added at the beginning of each subsection.

2.5.1 Insight on research question 1

RQ: What are the main open challenges and emerging research trends in supervised and unsupervised person re-identification, according to the latest literature?

Summarizing the key insights from Section 2.2 and Section 2.3, several points are touched in this chapter involving this research question.

In the supervised context, two major research directions — feature learning and metric learning — dominate the field. The former focuses on improving neural network architectures to obtain discriminative embeddings robust to variations in appearance. Recent trends include the adoption of ResNet [22] and ViT [23] that incorporate in its architecture global and local features. The latter, metric learning, advances through the design of improved loss functions to better guide representation learning. The combination of classification and triplet losses has shown strong results [16], [19]. More recent approaches, such as CLIP-based methods [123] and mask-guided attention models [30], currently represent the state of the art in supervised Person ReID.

Section 2.3 discussed an additional challenge that arises in the unsupervised setting: learning from noisy pseudo-labels. In the vanilla algorithm (Figure 2.5), noise is introduced during clustering, sampling from memory banks, and network parameter updates. To mitigate such noise, teacher-student frameworks [141], momentum-based strategies [163], [164], and more recently, self-paced learning approaches [148] have been proposed.

Overall, the main challenge in both supervised and unsupervised settings is developing models that generalize across individuals, as person re-identification inherently requires adapting to different identities between training and testing sets. Data quality and quantity are crucial, as large, accurately labeled datasets are difficult to obtain. Transferring knowledge from unlabeled sources to controlled scenarios represents a promising direction. Addressing these challenges will enable models robust to appearance variations and consistent performance across domains.

2.5.2 Insight on research question 2

RQ: How can embeddings for person re-identification be extracted to capture only identity-specific features, minimizing influence from transient attributes?

In this chapter we saw that existing person re-identification methods largely emphasize discriminative power rather than identity-specific representation. This occurs because discriminative models are easier to optimize and naturally learn features correlated with visual appearance cues in the dataset.

However, identity-specific features operate on a domain with no fixed parameters: Currently, there is no established metric that directly measures how well a model captures identity itself, independently of transient appearance cues. Consider, for instance, the assumption of ReID in short- and long-range scenarios and the role of clothing [13]: clothes may serve as an effective cue in short-term settings but become transient attributes in the long-term. For instance, the datasets we considered in our study, namely Market-1501 [3] and DukeMTMC [4], inherently rely on the short-term assumption for clothing. Despite this limitation, the proposed techniques in [14] can still exploit cues such as bags, caps, or even the presence of a bicycle, allowing re-identification models to operate at a finer level of granularity within the available constraints.

Nevertheless, this reliance on transient cues is reinforced by the typical use of pre-trained backbones (e.g. [19], [123] ImageNet-based), which tend to emphasize object-level semantics such as texture, color, and local patterns, rather than identity-consistent representations because of the filter the network learned.

Anticipating what is discussed in the next chapter, we addressed this direction in Chapter 2 through diffusion models [98], aiming to decouple the core embeddings from non-salient information. This generative approach offers a potential line of investigation toward learning identity-specific representations.

2.5.3 Insight on research question 3

RQ: How robust are current state-of-the-art person re-identification models to domain shifts, and what is the effectiveness of domain adaptation strategies in bridging performance gaps?

The proposed UDA strategies in Section 2.3 highlighted their effectiveness of domain adaptation strategies and demonstrated to be the most effective research areas in person re-identification. In the chapter several approach as presented to mitigate domain shifts such as the curriculum learning paradigm [77]. Despite these successes, the robustness remains uncertain and for this, we try to quantify such robustness on the proof-of-concept proposed in Chapter 4.

Chapter 3

PoC 1: Generation and Manipulation of embeddings

The work of this chapter moves under the umbrella techniques of Generative and especially the Diffusion models. It is published in Sensors [98] with the title *A Generative Approach to Person Re-identification*.

Generative models class had extensive usages in person re-identification tasks, such as data augmentation ([180], [181]), Many state-of-the-art methods have been proposed to use generative techniques for this purpose, as seen in [182]: by varying poses and viewpoints, these models generate diverse training samples, allowing the training of more robust re-ID models. Other common applications include cross-domain adaptation ([183]–[185]) and feature disentanglement ([186]).

Once trained to generate instances of a class, generative models encode each sample as a point in a latent space: a mathematical space where each point corresponds to a compressed multi-dimensional representation of an input instance, capturing its most relevant features (the semantic attributes) ([187]). Each latent point is typically represented as a vector of numbers, where each number has a specific meaning for the model. Person re-identification or similarity computation can then be performed by comparing these vectors in the latent space ([186]).

This was also the background of the idea the paper discussed in this chapter: [98]. In the case of humans, such latent embeddings could carry distinctive features useful for re-identification. [182] validates the principle that explicitly generating variance in non-identity factors (pose, view) drives the model to learn invariant identity representations. Our goal was therefore to investigate whether this property also holds for diffusion models [96], [97]—a recent state-of-the-art generative technique.

In [98] the proposed exploration analysis was something not yet done in literature at the moment of the publication and this led towards a *proof-of-concept* of two conceptual questions regarding the diffusion models and Person ReID:

1. Is it possible to characterize a person’s identity (from the person ReID perspective) as the conditional information required to generate a diversified set of images relative to the given person?
2. In that case, is it possible to compute a right inverse to the generator, extracting from an image of a person their conditioning identity?

Although from the point of view of person re-identification our results are slightly below the state-of-the-art, our work provides evidence that both questions are likely to have a positive answer. By conditioning generation on a fixed identity, we impose a structural constraint within the latent embedding that isolates invariant features. The inversion model (right inverse) then exploits this structure to disentangle the identity vector from the instance-specific noise.

3.1 Denoising Diffusion Models

In this section, we provide a concise overview of diffusion models. Readers familiar with this area may opt to skip this section.

Denoising Diffusion Models (DDM) [96] mark a significant breakthrough in deep generative modeling, posing a challenge to the longstanding predom-

inance of Generative Adversarial Networks [188]. These models have been applied with great success in various recent and notable projects, such as [189]–[191]. Key factors contributing to their rising prominence include unparalleled generation quality, high adaptability and ease of conditioning, the ability to generate diverse and robust samples, stable training processes, and remarkable scalability.

An important characteristic of this generative paradigm is its robust probabilistic foundation. However, a formal exploration of the underlying theory behind denoising diffusion models is beyond the scope of this article. For a deeper theoretical understanding, we refer the reader to the extensive literature available [96], [97]. Here, we aim to provide an operationally focused description of the denoising model, comprehensible even without a deep understanding of its theoretical background.

In very rough terms, a diffusion model trains a single network to denoise images with a parametric amount of noise. This network is then used to generate new samples in an iterative manner, starting from a purely ‘noisy’ image and progressively removing a decreasing amount of noise.

This process is traditionally referred to as *reverse diffusion*, as it aims to “invert” the *direct diffusion* process, which consists of iteratively adding noise (see Figure 3.1).

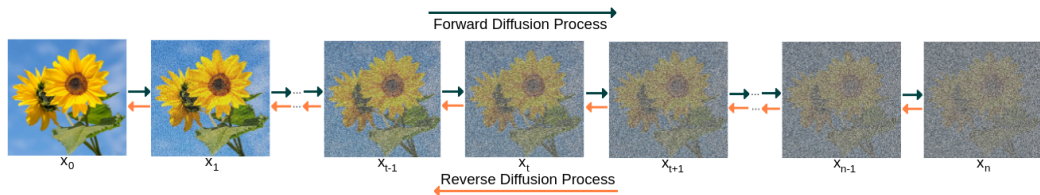


Figure 3.1: Direct and Reverse diffusion. Picture from [192]

We concentrated on a particular subclass of diffusion models, known as Implicit Diffusion Models [97], which have been a focal point of our research in numerous previous works. A key feature of these models is their fully deterministic reverse diffusion process. This attribute is crucial for applications

that require embedding the output back into its latent representation, as discussed in [193]. Another significant advantage of Implicit Diffusion Models is their efficiency, typically requiring only a minimal number of iterations (about 10, as reported in [187], [194]) — a stark contrast to other techniques that may need thousands of iterations. Additionally, their remarkable conditioning capabilities have been demonstrated in the field of precipitation forecasting [192], where the goal is to predict the probability distribution of a specific atmospheric parameter based on recent weather conditions.

3.1.1 The denoising network

The only trainable component of the reverse diffusion process is a *denoising network* $\epsilon_\theta(x_t, \alpha_t, c)$, which takes as input a noisy image x_t , a signal rate α_t , possibly a condition c , and tries to guess the noise present in the image. The model is trained in a completely supervised way. The main steps are the following:

1. pick a random image x_0 from the train set, coherent with the condition c ;
2. select a random step t of the process; to each step t is associated a signal rate α_t defined by a suitable noise scheduling (more in the sampling section);
3. sample a random gaussian noise ϵ ;
4. create a corrupted image as a weighted combination of x_0 and ϵ :

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon$$

5. train the network to properly guess the amount of noise present in x_t , by minimizing the distance between $\epsilon(x_t, \alpha_t, c)$ and ϵ

The previous steps are summarized in the pseudocode of Algorithm 1, where $q(x_0|c)$ denotes the distribution of training data, relative to the condition c .

Algorithm 1 Training

```

1: repeat
2:    $x_0 \sim q(x_0|c)$  ▷ take a sample coherent with  $c$ 
3:    $t \sim \text{Uniform}(1, \dots, T)$  ▷ choose a timestep
4:    $\epsilon \sim \mathcal{N}(0; I)$  ▷ create random gaussian noise
5:    $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon$  ▷ corrupt the sample with signal rate  $\alpha_t$ 
6:   Take a gradient descent step on  $\|\epsilon - \epsilon_\theta(x_t, \alpha_t, c)\|^2$  ▷ backpropagate the loss
7: until converged

```

The approach to conditional training previously described was formally examined in [195]. The paper proposes a blend of conditional and unconditional generation, a method some researchers found beneficial for training, though we observed no significant improvements.

Another strategy for implementing conditioning involves using an auxiliary classifier [188], akin in concept to AC-GANs [196]. This method trains a classifier $f_\phi(c|x_t)$ on a noisy image x_t to predict its class c . The gradient $\nabla_x \log f_\phi(c|x_t)$ is then utilized to steer the diffusion sampling process, enhancing the generation of samples that more accurately reflect the given condition. This technique remains untested in our work but represents a potential avenue for future enhancements.

The preferred architecture for the denoising network adopts a U-Net structure [197]. Our specific version is discussed in Section 3.2.2.

To heighten the network’s sensitivity to the signal rate α_t , this parameter is often encoded through a bespoke sinusoidal transformation, dividing it into a series of frequencies, mirroring the technique of positional encodings in Transformers [198].

3.1.2 Sampling

Sampling is an iterative process. Starting from a purely noisy image x_T , we progressively remove noise by calling the denoising network. Specifically,

if the error predicted by the network at timestep t is $\epsilon = \epsilon(x_t, \alpha_t, c)$, then the corresponding denoised prediction is

$$\hat{x}_0 = \frac{1}{\sqrt{\alpha_t}}(x_t - \sqrt{1 - \alpha_t}\epsilon)$$

Next, ϵ is re-injected into the network at signal rate α_{t-1} , to get the next noisy image

$$x_{t-1} = \sqrt{\alpha_{t-1}}\hat{x}_0 + \sqrt{1 - \alpha_{t-1}}\epsilon$$

The pseudocode of the sampling process is given in Algorithm 2

Algorithm 2 Sampling

- 1: $x_T \sim \mathcal{N}(0, I)$
 - 2: **for** $t = T, \dots, 1$ **do**
 - 3: $\epsilon = \epsilon_\theta(x_t, \alpha_t, c)$ ▷ predict noise
 - 4: $\tilde{x}_0 = \frac{1}{\sqrt{\alpha_t}}(x_t - \sqrt{1 - \alpha_t}\epsilon)$ ▷ compute denoised result
 - 5: $x_{t-1} = \sqrt{\alpha_{t-1}}\tilde{x}_0 + \sqrt{1 - \alpha_{t-1}}\epsilon$ ▷ re-inject noise at rate α_{t-1}
 - 6: **end for**
-

A significant component of the reverse diffusion process is the scheduling of diffusion noise $\alpha_{t=1}^T$. The authors in [96] suggested the use of linear or quadratic schedules. However, such choices lead to a rapid decrease in the initial steps, creating challenges in the generation phase. The literature offers alternative scheduling functions that decrease more gradually, including the “cosine” or “continuous cosine” schedule [199], [200]. Opting for a gentler scheduler not only solves the generation issues but also reduces the number of iterations needed. For our purposes, we have implemented a generator with 10 diffusion steps.

3.2 The proposed Approach

Contrary to all such methods of the previous chapter, a distinct approach was adopted in [98] by associating a unique latent representation with each

identity. This representation was learned as the conditioning information necessary to generate diversified samples of the particular person from random Gaussian noise (see Figure 3.2a). The generator takes as input a (learned) embedding of the identity $emb(id)$, a noise vector z , and produces a sample image $GEN(emb(id), z)$. The next step involves computing a right inverse GEN^{-1} to the generator, producing, from an individual’s image X , an embedded identity id_X and a noise vector z_X , such that $GEN(id_X, z_X) = X$ (see Figure 3.2b). The idea is that id_X should capture the interesting and individual features of the person represented in X , while the “noise” z'_X should contain all information unrelated to the specific identity, such as background, pose, etc. The identity id_X is then used to retrieve samples from the gallery by comparing their embeddings in the latent space.

The inversion of the generator was the most delicate part of the work and is discussed in Section 3.2.1.

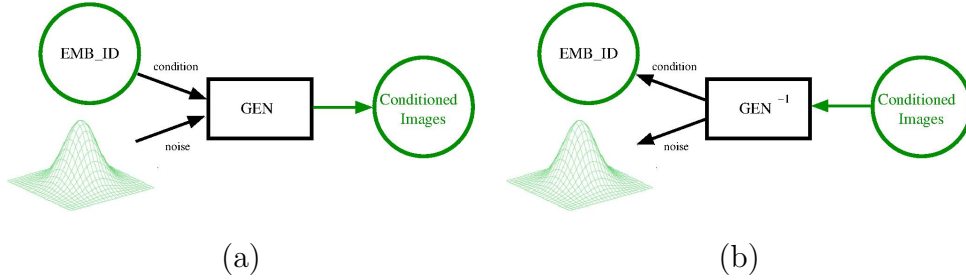


Figure 3.2: Conditional generative model and its inversion. We trained a diffusion generative model conditioned on an embedding of people’s identity. The embedding of identity labels was learned along with the generator. In reverting the conditional generator, we obtained a model that was able to split the image into the identity of the person, and all of the contextual information (pose, background, an so on) that was mapped back to the noise.

As generative model, we employed an instance of the recent diffusion models [96], [97] due to their remarkable sensitivity to conditioning, unbiased sampling capabilities and excellent expressiveness. Examples of generation of images conditioned on the individual identity are given in Figure 3.3.



Figure 3.3: Conditional generation. On the left, three sample identities taken from the Market1501 dataset. We sample 5 random noises, and for each of them we generate an image conditioned by the given identity; the noise in each column is always the same.

The generative quality is not perfect, and can probably be improved upon. In particular, for this prototyping work, we used a low resolution version of the dataset, with images of a 64×32 dimension (instead of the usual 128×64). The embedding space for the identities has 32 dimensions. In spite of being so small, the conditioning seems to work well, as exemplified in Figure 3.3.

The results presented in this article represent a preliminary investigation into the direction described above. They are primarily intended as a proof-of-concept, testing the feasibility of the approach. The current performance, measured on standard benchmarks such as Market-1501, is encouraging, though it falls short of various state-of-the-art approaches.

3.2.1 Methodology

As introduced earlier, the core concept is to assign a unique latent representation to each identity, which is learned as the necessary conditioning information for generating specific individual samples from random Gaussian noise. The overall structure of the generative model is depicted in Figure 3.2a. This model is an example of a diffusion model, as discussed in the preceding section.

The subsequent step involves reversing the generative model, as shown in Figure 3.2b. This process enables us to derive, from an individual’s image, the latent representation of their identity and noise that captures all details unrelated to identity, such as background and pose. The latent identity holds significance for person re-identification because comparing it with the latent representations of gallery samples allows us to identify the most similar identities.

The feasibility of inverting diffusion models through appropriately trained neural networks has been explored in many of our prior works. The overall idea was presented in [193]; this approach was applied in [187] for the “reification” of artistic portraits by embedding a portrait into a latent space of human faces and reconstructing the closest real approximation. In [194], the diffusion inversion was utilized to create a trajectory in the latent space that induces a smooth rotation effect on human faces.

In this work, after several trials, we decided to divide the inversion networks into two separate models, as illustrated in Figure 3.4. One model, named *Img2ID*, converts images into latent identities, while the other, named *Img2Noise*, generates the noise. Given that both models extract orthogonal features from the source image, there was no compelling justification for them to share layers.

There are a few different possibilities for training the inversion models. Suppose to have a labelled image $\langle id, X_{id} \rangle$. We can train *Img2ID* to minimize the distance

$$\|emb(id), img2ID(X_{id})\| \quad (3.1)$$

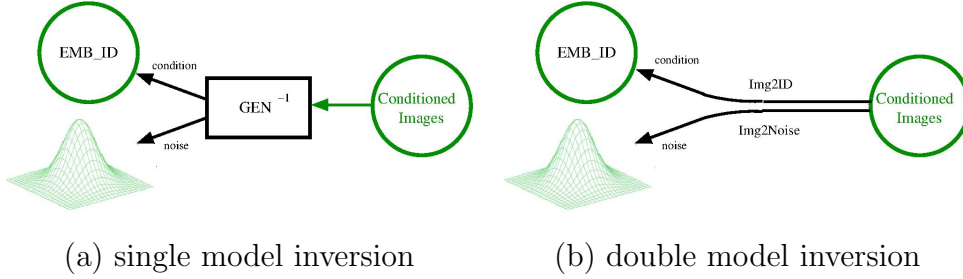


Figure 3.4: The inversion of the generator can be realised through a single network (left), or split in two distinct networks.

where emb is the identity embedder of the generator. Similarly, $Img2Noise$ can be trained to minimize the error

$$\|GEN(Img2Noise(X_{id}), emb(id)), X_{id}\| \quad (3.2)$$

as the noise synthesized by $Img2Noise(X_{id})$ should result in the generation of X_{id} when conditioned over $emb(id)$. One issue with this approach is that the Market1501 dataset is relatively small. When constraining the inversion to images from the training set, where identities are disjoint from the gallery, the technique struggles to generalize effectively to new identities.

Given that we have a generator at our disposal, a natural idea is to leverage it to expand the training data. We select a random identity from the latent space of identities, along with a random noise ϵ , and generate an image $X_{id} = Gen(\epsilon, id)$. We then proceed as previously described. However, a challenge arises from not knowing the distribution of identities in the latent space, which could be arbitrary, making it difficult to select a random identity. Our solution involves utilizing an auxiliary id-generator to learn the distribution of identities in the latent space, enabling us to sample effectively within this space.

Results of the inversion process are shown in Figure 3.5. The reconstruction is satisfactory, although not flawless. Some details are lost, which could pose challenges for Person Re-ID.

For those interested, we also display the noise extracted from the image, which ideally contains all information not specific to the individual (middle

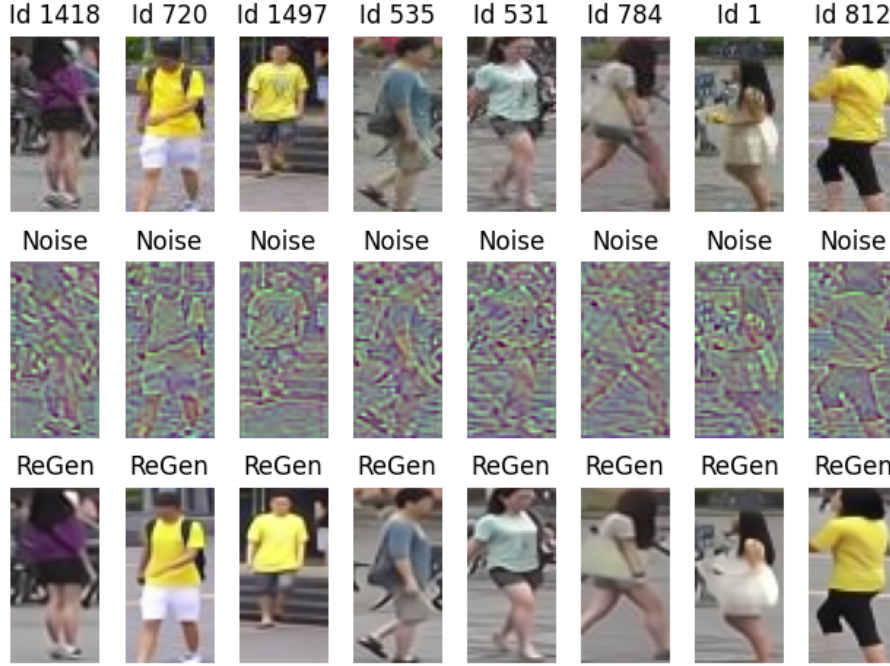


Figure 3.5: Reconstruction examples. In the first row we have original gallery images, in the middle row the noise synthesized by inversion, and in the third row the reconstructed images.

row). The noise, assumed to have a Gaussian distribution, has been clipped within the range $[-2.5, 2.5]$ and then renormalized to $[0, 1]$ before visualization.

The inverted generator obtained through this process is extremely robust. In principle, it enables the derivation of the internal representation of any individual within any context, provided that the individuals and contexts are akin to those encountered in the training set.

3.2.2 Neural Networks architectures

In Section 3.1, it was clarified that the only trainable component within a diffusion model is the denoising network $\epsilon_\theta(x_t, \alpha_t, c)$. This network receives inputs including a noise rate α_t , an image x_t corrupted with a corresponding

level of noise, and a condition c , which in our scenario represents the identifier of an instance. The objective is to predict the noise present in the image.

This task aligns with the conventional image-to-image paradigm, and the denoising network is typically constructed using the well-established U-net architecture [197]. The architecture of our implemented denoising network is illustrated in Figure 3.6; the main constituent blocks are depicted in Figure 3.7.

As is customary for U-nets, our network follows an encoder-decoder structure with skip connections that link corresponding layers in the encoder and decoder. This architectural choice facilitates the assimilation of both global and local structures within the image, making it highly compatible with the demands of the diffusion model.

The U-Net typically relies on parameters such as the number of downsampling blocks and the number of channels for each block, with the upsampling structure being symmetric. In our implementation, we opted for four downsampling blocks with respective channel dimensions [48, 96, 192, 384].

We emphasize that our work deals with images at a resolution of 64×32 ; for higher-resolution images such as 128×64 , an additional layer in the U-chain may be necessary.

The input pertaining to the noise variance α_t is incorporated using sinusoidal positional embeddings. Subsequently, this information is vectorized, concatenated, and forwarded through the blocks along with the conditioning information of the id and the processing of the noisy image.

Regarding the inversion networks, Img2Noise entails an image-to-image process, as the noise maintains the same spatial dimension as the input image. Based on prior studies [193], the U-Net architecture has demonstrated superior performance for such tasks.

The Img2Id model follows a similar approach, employing only the encoder portion of the U-Net. The processing is then finalized through a compact sequence of dense layers.

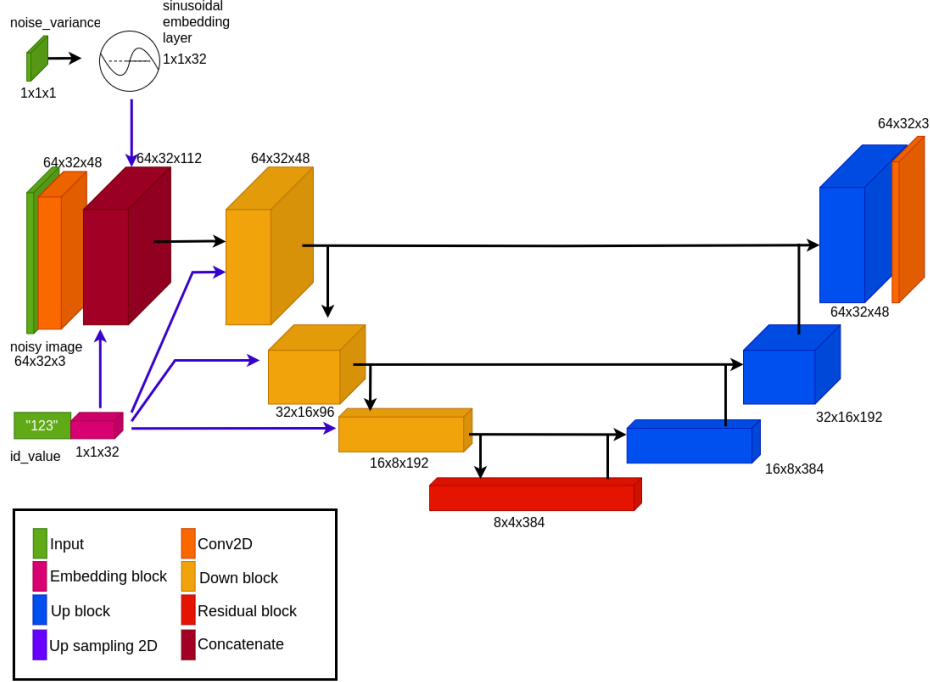
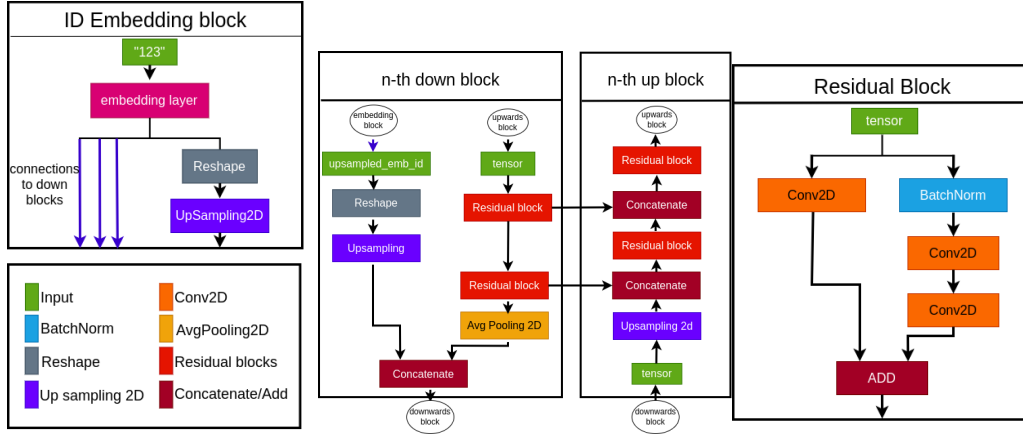


Figure 3.6: Denosing Network Architecture.

3.3 Evaluation

We conducted experiments utilizing the already introduced Market-1501 dataset [3]. The current performance of our technique falls somewhat below the state-of-the-art. We achieved a mean average precision (mAP) of 73%, whereas state-of-the-art techniques attain a MAP of 96% in the supervised setting and approximately 90% in the unsupervised setting. The mAP score was computed following the guidelines provided by the dataset.

Visually, the results are satisfactory, and the errors made by the model are quite understandable given the low resolution we are working with. Distinguishing between certain identities at this resolution presents a significant challenge (see Figure 3.8).



(a) ID embedding block (b) N-th down and up blocks (c) Residual block

Figure 3.7: Main blocks of the Denoising Network. Figure 3.7a: the input ID undergoes processing through an embedding layer; the resulting output, after reshaping and upsampling, is concatenated with the various downblocks. Figure 3.7b: the n th downblock receives two inputs: the embedded id and the output of the previous block. The block outputs two elements: the skipping connections for the upblock at the same level and the output for the next block. Figure 3.7b: a traditional Residual Block.

3.3.1 Latent space exploration

In our approach, latent representations are synthesized as the features of each individual able to shape the generation of distinctive images of the person under different noises. As is typical in generative processes, this naturally leads to the creation of a highly informed and well-structured latent space, where similar persons yield similar encodings.

Figure 3.9 presents some examples: in the first column, we have several identities, and in the corresponding rows, we display five different individuals whose latent representations are closest, in terms of Euclidean distance, to the representation of the given identity. These images are sourced from the traditional Market1501 dataset [3], which we primarily used for our experiments. Notably, we observe the significant similarity between different identities, underscoring the non-trivial nature of the reidentification problem.



Figure 3.8: Person re-identification. In each row, the first image is the query, and the successive images are the 5 best matchings in the gallery.

In Figure 3.10 we show the two closest identities, at a minimum distance of .14, and the two more apart from each other, at a maximum distance of .99. The average euclidean distance is around .6.

Another intriguing operation involves deciphering how different explanatory factors of variation in the data are captured within the latent representation, akin to understanding the meaning of nucleotide sequences in a genome. However, these factors can often be entangled, meaning that a visible effect may depend on a combination of latent variables rather than a single one.

Our studies in this direction are very preliminary. As an illustration, in Figure 3.11, we demonstrate the effect on generation when modifying a single variable in the latent space of identities.



Figure 3.9: Inherent clustering of the latent space. For every id in the initial column, we display five distinct individuals whose latent representations are closer, in terms of Euclidean distance, to the representation of the specified id.

It’s noteworthy to observe how the modification of a single variable can correspond to a significantly different identity.

Furthermore, investigating interesting semantic trajectories within the latent space can also facilitate the synthesis of new and distinct identities, which can be valuable, for instance, for data augmentation purposes.

3.3.2 Ablation and alternatives

Several iterations of the aforementioned architecture have undergone testing. Prior to delving into learning the distribution of latent variables using an auxiliary generator, we experimented with various regularization techniques to shape its distribution. This ranged from straightforward methods such as employing BatchNormalization layers to injecting a mild amount of noise.



Figure 3.10: On the left we have the two closest identities in the latent space of Market1501; on the right, the two more distant ones.

Subsequently, we explored a comprehensive variational model, that should enforce the adoption of a Gaussian-shaped latent space through Kullback-Leibler regularization.

However, a drawback of the latter approach is the relatively weak signal influencing the definition of latent representations. This poses a challenge in balancing it with the KL-component, as highlighted in [201]: in such a situation, the KL-component could dominate, leading to the well-known variable collapse phenomenon [202].

The utilization of an auxiliary generator can essentially be conceptualized as a two-stage generative model [201], [203]. Initially, we experimented with a variational autoencoder for implementing the auxiliary generator. However, this endeavor resulted in a notable increase in reconstruction loss, which corresponded to a loss of variance in generated data [204]. In contrast, a diffusion model appears to learn the distribution much more effectively in this regard.

Exploring a different research direction, we sought to increase the separation between distinct latent representations through the incorporation of a "repulsive" loss, which was proportionate to the inverse of their distance in the space, akin to a form of magnetic repulsion. This was complemented by a weak attractive loss, which encouraged points to remain close to the origin,

effectively centering the latent space around it.

While we retained the latter, the repulsive loss did not appear to enhance performance and could potentially be detrimental. Our hypothesis is that the loss primarily impacts latent variables with minimal significance, attempting to segregate them across different subjects. This segregation is artificial and lacks semantic predictability, rendering it unreconstructable from the input images.

Finally, we conducted several experiments on the Embedding module for identities. In the current implementation, it comprises a single Keras Embedding layer. Initially, this layer expands labels to their categorical encoding and subsequently maps them to a latent space of the desired dimension through a single dense layer.

In principle, this setup should not impose any restrictions, as any embedding can potentially be learned from a categorical description. However, what remains unclear is the model’s flexibility to adapt and enhance encodings during training, along with the acquisition of additional knowledge. To explore this aspect, we experimented with more complex embedding modules, although they did not yield any discernible benefits.

3.3.3 Conclusions

The work presented in this article is first of all a *prof-of-concept*: we aimed to test the conditional capabilities of generative diffusion models in a complex scenario. Specifically, we sought to learn the latent representation of different individuals as the shared information necessary to condition the generation of images of the given person from varying noise. This approach enables the separation of an individual’s identity from other instance-specific information, such as pose and background, expressed as part of the noise.

There are numerous avenues for improvement in this work. Both the generator and the network used for its inversion could be enhanced. A natural research direction involves repeating the experiments with images at the original 128x68 spatial resolution, potentially expanding the dimension of

the latent space. Additionally, a detailed analysis of the errors made during the re-identification process could lead to valuable insights and potential enhancements. Leveraging the generator for intelligent and targeted data augmentation presents another intriguing possibility.

A limitation of our approach is its reliance on supervised training: to train the conditional generator, a significant number of instances of the same identity in different poses and contexts are required. Exploring whether the methodology can be extended to an unsupervised scenario presents a complex and interesting challenge. It is plausible that typical unsupervised Person ReID approaches, predominantly based on clustering, centroids, and pseudo-labels [183]–[185], could be adapted to our approach.

3.4 Conclusions on Research Questions

3.4.1 Insight on research question 2

RQ: How can embeddings for person re-identification be extracted to capture only identity-specific features, minimizing influence from transient attributes?

In contrast, with traditional metric-learning approaches, the generative perspective explored in this work [98] allows a clearer separation between stable identity features and transient noise.

The experiments and comparative analysis in this chapter provided insights and proof-of-concept of the proposed approach. See Figure 3.4 for reference. By decomposing the latent representation into two components: one associated with the subject’s intrinsic identity and the other capturing transient factors; The proposed method explicitly models what should remain invariant across appearances and thus, the core embedding is isolated with the identity-specific parts. See also Figure 3.5

In summary, embeddings generated under this paradigm capture in theory more intrinsic, person - specific cues and are less sensitive to environ-

mental or contextual variations. This confirms that modeling and removing transient components—rather than relying solely on metric discrimination—is a promising direction for achieving identity-consistent representations in person re-identification that worth a line of investigation.

Although the results are promising, further validation on larger and more diverse datasets is needed to fully assess the generalizability and practical impact of this approach.

3.4.2 Insight on research question 3

RQ: How robust are current state-of-the-art person re-identification models to domain shifts, and what is the effectiveness of domain adaptation strategies in bridging performance gaps?

Despite no strictly response can be provided to this question by this work, we argue that generative models can simulate diverse domain conditions and significantly advance research in the context of domain adaptation techniques. Identity-conditioned generative models, such as diffusion-based approaches, offer a promising path to bridge this gap by creating controlled cross-domain variability. Yet, synthetic data often fail to reproduce the unpredictable anomalies of real environments. Identifying and modeling such anomalies through generative methods remains a key open-challenge step toward realistic datasets and, consequently, more robust domain adaptation.



Figure 3.11: Effect of latent variables on generation. In each row we keep a same random noise and vary a given variable in a predefined range. The specific variable and its value are indicated above the generated image.

Chapter 4

PoC 2: Domain Adaptation experiments

This chapter presents a *proof-of-concept* study on the domain adaptation problem in Person Re-Identification [99]. The domain gap arises from two main factors: the intrinsic variability of ReID scenarios, such as lighting, camera angle, viewpoint, and clothing differences, which makes models sensitive to environmental changes [3]; and the high cost of collecting and annotating person ReID data, which limits dataset diversity and reduces generalization [1], [3], [27].

The study describes a systematic pipeline of experiments to quantify how domain shifts affect model performance and assess strategies for improving generalization. Even high-performing models suffer significant drops when tested on unseen datasets [26], [205].

We found that Person ReID research has explored transfer learning, data augmentation, and domain translation or adaptation [26], [206], [207], yet few works systematically quantify the domain gap itself. This lack of benchmarks hampers reproducibility and fair comparisons.

The contributions of this work are twofold:

- We quantify the gains of expanding training sets with diverse, non-domain-specific data, evaluating trade-offs between large-scale augmen-

tation and lighter domain adaptation methods, such as one-shot or unsupervised learning, to provide actionable insights for future research

- A comparative analysis of CLIP-ReID models[120], [123] and strong image-only baselines, clarifying the benefits of multimodal pre-training in mitigating domain discrepancies and improving robustness in real-world deployment

In this work two CLIP-ReID models with different backbones are compared to their image-only baselines [123], directly assessing the contribution of multimodal pre-training. We also explore enlarging training data with non-domain-specific samples. The best-performing model (ViT CLIP-ReID [123]) is trained on a combined Market-1501 and Airport dataset, with and without Random Erasing Augmentation [19]. Results show an average improvement of +4.3% in mAP and +4.0% in Rank-1 accuracy, though diminishing returns appear as more datasets are added—highlighting that dataset selection matters more than volume.

4.1 Technical Background

In this article, we provide quantitative measures of the domain gap, focusing on recent CLIP-based models, which show promise for domain adaptation by aligning visual and textual representations in its original form [120]. Several models were evaluated on standard benchmarks, including Airport [1], ENTIRE-ID [5], Market-1501 [3], DukeMTMC-reID [4], and CUHK03 [25], training on one dataset and testing on others to measure cross-domain generalization. We invite the reader to take a look to the dataset and metrics in both sections: Section 1.1.3 and Section 1.1.4 in Chapter 1.

Next section, Section 4.1.1 describes the class of models under investigation and the data augmentation techniques employed for the experiments.

4.1.1 Models

We focused on the CLIP-ReID framework [123], which is the adaptation of the CLIP framework [120] by OpenAI to person re-id. This is due to a number of factors.

First and foremost, person re-id is conceptually similar to zero-shot classification [120]: models will be evaluated on (and faced with once they are deployed in a real-world scenario) a set of identities that were not present in the training set. If the identity of each person is considered to be a classification label, then the person re-id task becomes the classification of unseen labels (zero-shot classification [120]) taken from an open-ended set of labels, which CLIP excels at. Additionally, CLIP models are able to learn fine-grained visual features which, together with the nearly unlimited scope of natural language that CLIP is able to exploit, allow it to easily adapt to other tasks and, more importantly, domains. This ability to adapt to new domains should, hopefully, transfer to person re-id and help bridge the domain gap.

Another reason for selecting CLIP-Based models is their ability to learn rich and fine-grained visual representations which make it particularly suitable for person re-id, where subtle differences in appearance, pose, and context can be crucial for accurate identification.

The CLIP-ReID models also obtained the highest scores at the time of writing on the MSMT17 dataset [26] (one of the most challenging supervised person re-id datasets) on the Papers with Code website.

Together with the CLIP-ReID models, we used the corresponding baseline models (which consist of pure CLIP’s image encoder [123]) to measure the impact of the CLIP-ReID framework in the context of the domain gap. In this section we introduce the CLIP framework (Section 4.1.1), the baseline models (Section 4.1.1) and the CLIP-ReID framework (Section 4.1.1).

Clip

CLIP (Contrastive Language-Image Pre-training) [120] is a framework by OpenAI which aims at associating images and natural text. This allows the models to exploit natural language features which have a nearly unlimited scope. Specifically, given N images and N prompts, CLIP will determine which of the $N \times N$ possible $(image, text)$ pairings actually occurred. This is achieved by jointly training a text encoder (a transformer [208]) and an image encoder (either a ResNet [22] or a Vision Transformer [23]) on a shared embedding space, using a dataset of roughly 400 million $(image, text)$ pairs extracted from the internet.

In practice, CLIP learns a multi-modal embedding space, which is shared by both the image and the text features, and it maximizes the cosine similarity of the true N pairs of the batch while minimizing it for the remaining $N^2 - N$ negative pairs. In more mathematical terms, given a batch of size $B \in \mathbb{Z}^+$ composed of $\{img_1, \dots, img_B\}$ images and $\{text_1, \dots, text_B\}$ pieces of text, the similarity between image img_i and text $text_i$ is computed as [120], [123]:

$$S(V_i, T_i) = g_V(\mathcal{I}(x_i)) \cdot g_T(\mathcal{T}(text_i)) \quad (4.1)$$

Where:

- $\mathcal{I}(\cdot)$ and $\mathcal{T}(\cdot)$ are the functions computed by the image and text encoders, respectively.
- $g_V(\cdot)$ and $g_T(\cdot)$ are the linear layers that project the given (image and text, respectively) embedding into the shared embedding space.

These similarities are optimized using two contrastive losses, called image-to-text and text-to-image respectively [120], [123]:

$$\mathcal{L}_{i2t}(i) = -\log \frac{\exp(S(V_i, T_i))}{\sum_{a=1}^B \exp(S(V_i, T_a))} \quad (4.2)$$

$$\mathcal{L}_{t2i}(i) = -\log \frac{\exp(S(V_i, T_i))}{\sum_{a=1}^B \exp(S(V_a, T_i))} \quad (4.3)$$

In these two losses, the numerator uses the similarity of the *image, text* pair that actually matches, and thus has to be maximized, whereas the denominator uses the similarities between the other, non-matching, pairs, which will be minimized.

The immense scope of natural language allows CLIP to be easily adapted to other tasks, without the need for further training (which is referred to as zero-shot setting [120]). One important example of this is classification: CLIP can easily be adapted to classification by crafting a prompt for each label (e.g. "A photo of a {label}."), as can be seen in Figure 4.1.

In the context of classification, Zero-Shot CLIP has obtained results competitive with a fully trained baseline, outperforming it on 16 out of 27 datasets [120].

Baseline Models

We used Li, Sun, and Li’s baseline [123] architecture, which consists of CLIP’s image encoder. Specifically, we used a ResNet-50 and a ViT-B/16:

- For the **ResNet** model, the standard ResNet-50 architecture was enhanced by using the ResNet-D architecture [102], [120], the anti-aliased rect-2 BlurPool by Zhang [209] and by replacing the global average pooling layer with an *attention-pooling* layer (meaning a single layer of multi-head QKV attention, similar to a transformer mechanism, where the query is based on the globally average-pooled representation of the image) [120].
- For the **ViT** model, the CLIP’s implementation follows the original [23] closely, with the only addition of a normalization layer to the final combination of patch and position embeddings [120].

The Baseline models, pre-trained with CLIP and fine-tuned using standard person re-id losses (triplet loss and ID loss, Section 4.1.1), as seen

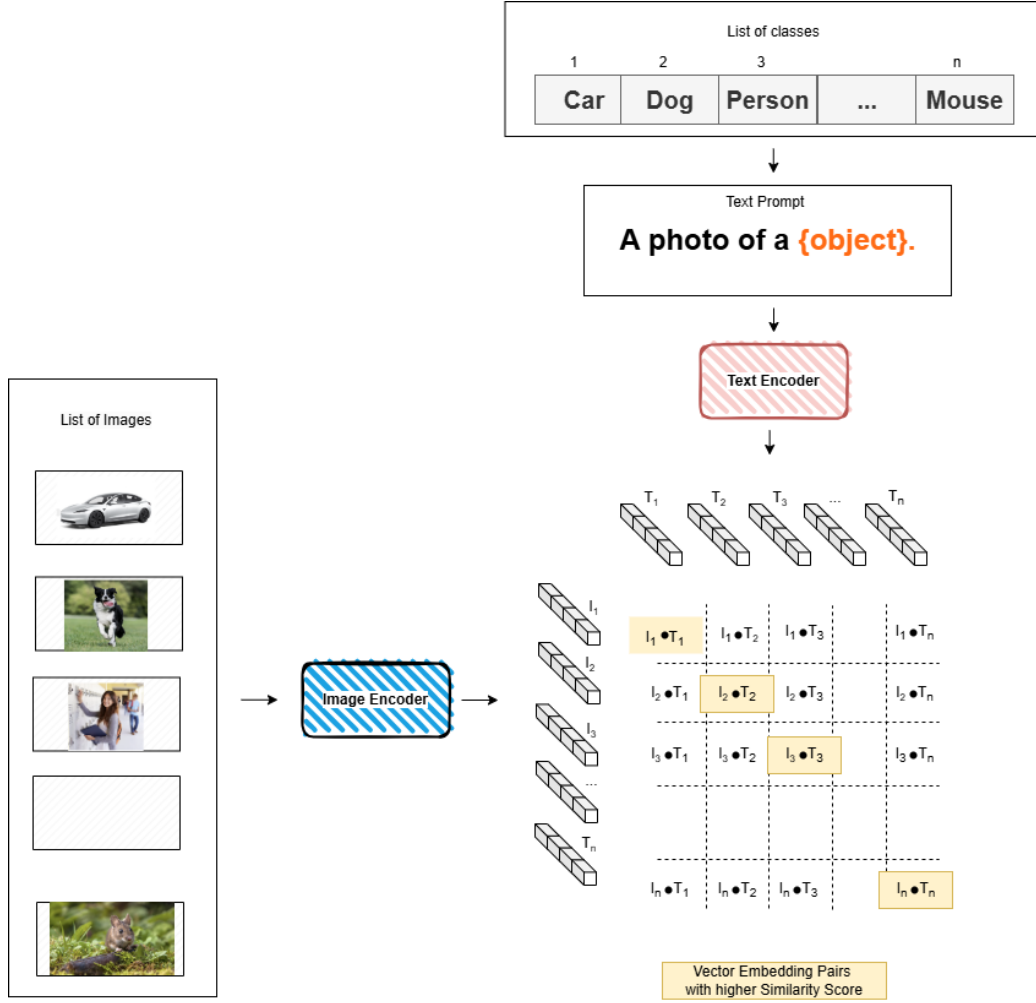


Figure 4.1: An example of CLIP being adapted to image classification.

in Figure 4.2, already show superior performance compared to other common baselines pre-trained on ImageNet (such as the Bag of Tricks Strong Baseline [19] for CNN-based models and the TransReID baseline [86] for ViT-based models) [123]. This suggests that CLIP’s pre-training imparts more robust and transferable features for person re-id than ImageNet, even without the additional enhancements of the full CLIP-ReID framework.

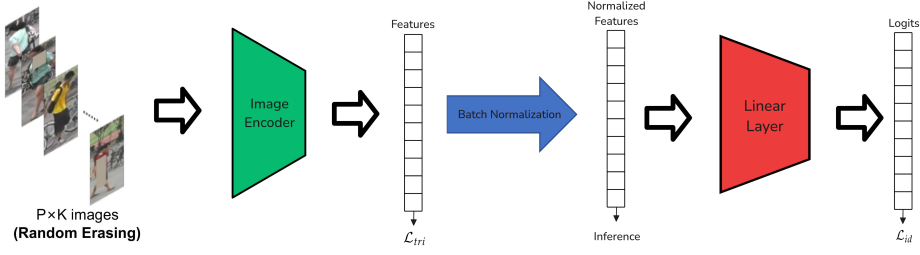


Figure 4.2: Architecture outline for the baseline models. $\mathcal{L}_{tri}$ is a triplet loss, and $\mathcal{L}_{id}$ is a classification loss for the different identities.

CLIP-ReID

Person re-identification is conceptually similar to zero-shot classification: if the identities of the people present in any given supervised dataset are seen as classification labels, then person re-id can be interpreted to be the classification of never-before seen identities (meaning, as said, a zero-shot classification problem). This similarity implies that CLIP, which excels at zero-shot classification, should adapt well to person re-id.

While exploiting CLIP seems like a promising premise, in practice it presents some challenges. Namely, the person labels used in person re-id are simple numerical values which lack any concrete natural language description. This has first been explored by Li, Sun, and Li [123]. The lack of exploitable meaning in the IDs was overcome by dividing the training in two stages:

- During the first stage, the two encoders are kept frozen (meaning their weights do not get altered), and a set of learnable tokens $[X]_1, \dots, [X]_M$ is optimized by using it to construct a textual prompt "A photo of a $[X]_1 \dots [X]_M$ person", as can be seen in Figure 4.3. This effectively overcomes the need for a concrete text label. The idea of using learnable tokens to create CLIP's textual prompt was first introduced by Zhou, Yang, Loy, *et al.* [210], who sought to overcome the limitations of manual prompt engineering. The loss function used during this stage is the sum of the two clip losses of Equation 4.2 and Equation 4.3.

- During the second stage, the learnable token’s weights are frozen and the image encoder’s weights are unfrozen, as can be seen in Figure 4.3. The training loss function used in the stage is a combination of ID loss, Triplet loss [19], [123], and an adaptation of the $\mathcal{L}_{i2t}$ loss of Equation 4.2.

Each image img_i in the training dataset is expected to have a person ID (pid for short) associated with it, y_i . This pid is used to compute the textual prompt (specifically the tokens) associated with image img_i as "An image of a $[X]_1 \dots [X]_M$ person."

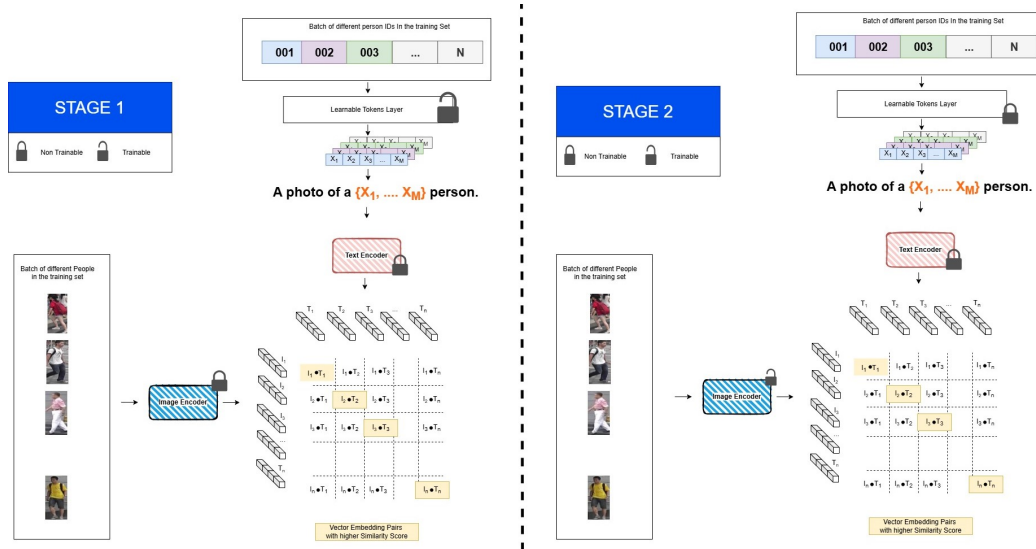


Figure 4.3: The CLIP-ReID framework [123]. The framework trains the model in two stages. In the first stage, embeddings for tokens representing each ID are learned and concatenated with the text prompt embedding, "a photo of a [...] person." This combined input is fed into the text transformer to calculate the loss. In the second stage, the visual encoder is trained, guided by embeddings from the frozen text encoder.

In Table 4.3 we provide the results of the CLIP-ReID framework computed by the authors on some of the datasets considered for our experiments. We mostly report them for the sake of completeness and for facilitating a more direct comparison.

Table 4.1: Scores obtained by the CLIP-ReID framework [123] and by the corresponding baseline architecture.

Backbone	Method	MSMT17		Market1501		DukeMTMC		Occ. Duke	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1
ResNet-50	Baseline	60.7	82.1	88.1	94.7	79.3	88.6	47.4	54.2
	CLIP-ReID	63.0	84.4	89.8	95.7	80.7	90.0	53.5	61.0
ViT-B/16	Baseline	66.1	84.4	86.4	93.3	80.0	88.8	53.5	60.8
	CLIP-ReID	73.4	88.7	89.6	95.5	82.5	90.0	59.5	67.1

As it is clear from the Table 4.1, the ClipReID framework appears to increase the model’s performance with respect to the Baseline across all datasets.

The scores can be improved with the embedding of side information (e.g. to make the model aware of the camera) and by using overlapping patches [123]. Re-ranking [9] (the practice of reordering the ranked list produced by the model, usually by searching the list using the first image as a query) can further improve the scores, which become 86.7 mAP and 91.1 Rank-1 on the MSMT17 dataset [26] (the top scores on the Papers with Code website at the time of writing).

4.1.2 Data Augmentation

The standard pipeline for person re-id generally includes the *Random Erasing Augmentation* (REA) [19]. As discussed above, pedestrian bounding boxes often contain some form of occlusion, which contributes to the complexity of the problem [211]. In order to push the model to adapt to partial occlusion, the Random Erasing Augmentation was proposed by Zhong, Zheng, Kang, *et al.* [212]. REA works as follows: each image will have a probability of undergoing REA equal to p_e , and those that do undergo REA have a randomly selected rectangle erased with random pixels, as can be seen in Figure 4.4. The effects of REA have been extensively studied by Luo,

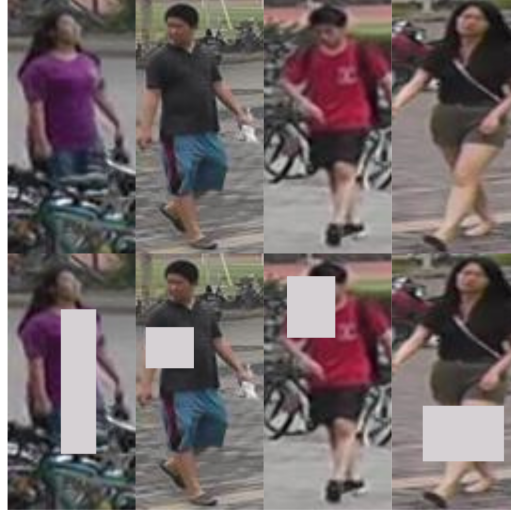


Figure 4.4: Examples of images undergoing REA.

Gu, Liao, *et al.* [19]. The study was conducted using a ResNet-50 [22] as a backbone and the Market-1501 [3] and the DukeMTMC-reID [4] datasets for training. During training each image was resized to 256x128 pixels, padded with 10 zero value pixels and then randomly cropped to a 256x128 image.

The images then underwent horizontal flipping with a 50% chance. Finally, each image was encoded in 32-bit float values in the range $[0, 1]$, and the RGB channels were normalized. The only training trick used in the baseline was a warmup period for the learning rate lasting 10 epochs. The impact of REA was measured in terms of mAP and CMC Rank-1 (Section 1.1.4) in two different settings: the *same dataset* setting, where the model was trained and tested on the same dataset, and the *cross-dataset* setting, where the model was trained on one of the two datasets and tested on the other.

Within the same domain, REA was found to improve both mAP and CMC Rank-1 by between 1% and 4%. However, in the cross-dataset setting, applying REA appears to harm the model’s performance by a similar margin. This is conceivably due to the model overfitting the training set in order to compensate for the extra occlusion [19].

4.2 Experiments results

In order to investigate the applicability of person re-id models into the real world, we sought to evaluate their adaptability to new scenarios which could differ widely from the conditions presented into any training dataset. Unsurprisingly, the great amount of variation that person re-id models have to account for cannot be fully captured by a single dataset [205] (both due to human appearance itself, the way this appearance is recorded and other environmental factors), and it has been found that models trained and evaluated on a dataset will perform poorly on data that differs noticeably from the training dataset, which is referred to as domain gap.

This domain gap is a known challenge in person re-id [26], [205] and other computer vision tasks [120] and is a fundamental obstacle to the deployment of a real-world re-identification system. Despite there being substantial research towards mitigating it, quantitative measures of its effects are not common. We therefore endeavored to accurately measure its effects on the performance of the CLIP-ReID models [123], and perform some small experiments attempting to mitigate them. Attempts towards bridging this gap have already been made, usually in the form of domain adaptation [26], [205] or achieving more generalized features, and although some success has been found, this challenge is yet to be fully overcome.

4.2.1 Cross-Dataset Tests

In order to assess the model’s performance on new data, we took the CLIP-ReID models (including the baselines) [123] trained on the Market-1501 [3] and DukeMTMC-reID [4] datasets, and measured their performance on different datasets. These tests are defined as Cross-Dataset Tests [206], [207], and are performed using the following datasets: CUHK03 [25], Airport [1], the test set of ENTIRE-ID [5], Market-1501 [3] (for models trained on DukeMTMC-reID) and DukeMTMC-reID [4] (for models trained on Market-1501). The results of these tests are shown in Table 4.2 and Table 4.3.

Table 4.2: Cross-dataset scores obtained by the CLIP-ReID models [123] and their baseline counterparts, all trained on **Market-1501** [3].

Backbone	Method	Airport		ENTIREID		DukeMTMC		CUHK03	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1
ResNet-50	Baseline	3.8	5.1	11.9	11.7	17.3	30.8	8.7	8.6
	CLIP-ReID	4.8	6.0	13.3	12.8	21.0	36.8	8.8	8.9
ViT-B/16	Baseline	16.0	18.4	29.3	29.2	44.8	64.3	34.9	36.9
	CLIP-ReID	20.0	22.4	38.9	38.4	50.2	68.9	38.6	40.4

The checkpoints trained on the DukeMTMC-reID [4] obtained similar cross-dataset scores, as shown in Table 4.3, but a bit lower than those obtained by their Market-1501-trained counterparts. This leads us to believe that market is, despite its smaller size, more complex than duke and thus models learn slightly more general features from it.

Table 4.3: Cross-dataset scores obtained by the CLIP-ReID models [123] and their baseline counterparts, all trained on **DukeMTMC-reID** [4].

Backbone	Method	Airport		ENTIREID		Market1501		CUHK03	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1
ResNet-50	Baseline	4.5	6.3	19.7	19.4	21.5	47.5	6.1	5.9
	CLIP-ReID	5.3	7.2	21.2	20.4	24.6	52.2	5.7	5.0
ViT-B/16	Baseline	10.0	12.9	33.6	32.9	37.4	62.2	19.2	20.4
	CLIP-ReID	14.3	17.4	40.2	40.3	43.4	70.7	25.9	27.6

When compared to the scores obtained by the same model in the traditional setting (Table 4.1), the degradation of performance is obvious. Across all models and all tests we performed, Airport [1] is the dataset where the lowest scores are obtained. This is probably due to the difference in scene between airport and most other datasets: both Market-1501 and DukeMTMC-reID were created from footage taken in a university campus [3], [4], with

the cameras being setup by researchers, whereas airport was created using the surveillance cameras of an actual airport. This leads to a wide difference in camera angles, image quality, activities, clothing etc., making Airport the most challenging dataset for models trained on Market-1501 [3] and DukeMTMC-reID [4].

4.2.2 Training on the Union of Datasets

In order to help close the domain gap, we re-trained the best performing model (ViT CLIP-ReID [123]) on the union of Market-1501 [3] (the original training dataset which yielded the best results in a cross-dataset setting) and Airport [1] (the dataset that proved to be the most challenging in a cross-dataset setting) in the hopes that more general re-id features could be learned from the resulting union. We first re-trained the ViT CLIPRe-ID model, using the same training algorithm and settings used by Li, Sun, and Li [123] and repeated the cross-dataset tests with the remaining datasets. The results can be seen in Table 4.4.

Table 4.4: Cross-dataset test result for the ViT CLIP-ReID model [123] trained on the union of Market-1501 [3] and Airport [1]. The difference in scores (Δ columns) are given with respect to the corresponding cross-dataset scores obtained by the same model trained on the Market-1501 dataset alone, as shown in Table 4.2. Scores for Market-1501 [3] and Airport [1] are also provided.

Dataset	mAP	Δ mAP	Rank-1	Δ Rank-1
ENTIRE-ID [5]	46.2	+7.3	45.9	+7.5
DukeMTMC [4]	53.5	+3.3	71.6	+2.7
CUHK03 [25]	41.8	+3.2	42.6	+2.2
Market-1501 [3]	89.0	-0.6	95.0	-0.4
Airport [1]	64.0	+44.0	56.9	+34.5

The idea of training a model on the union of two or more datasets is not

new. For example, Marchwica, Jamieson, and Siva [213] sought to achieve scene-independent re-id by training on a larger amount of data, obtained by merging between two and six datasets. It is worth mentioning that, following the work of Xiao, Li, Ouyang, *et al.* [214], they balanced the union of datasets by keeping ten images per person, ensuring to keep images from all cameras the person appeared in (except for when the person appeared in more than ten cameras, in which case one image per camera is kept). This is done to account for the fact that different datasets might have more or less images per person, which could lead to overfitting one of the datasets [213]. However, when merging only two datasets, this balancing can lead to a great reduction in the number of images used, so we opted to not balance the union.

Additionally, Luo, Gu, Liao, *et al.* [19] found that REA (Section 4.1.2), which is used in the CLIP-ReID training process [123], harms model performance in a cross-dataset setting, and therefore we repeated the above training process and testing procedures without it, the results can be seen in Table 4.5.

Table 4.5: Cross-dataset test result for the ViT CLIP-ReID model [123] trained on the union of Market-1501 [3] and Airport [1] **without using REA** (Section 4.1.2). The difference in scores (Δ columns) are given with respect to the corresponding cross-dataset scores obtained by the same model trained on the Market-1501 dataset alone, as shown in Table 4.2. Scores for Market-1501 [3] and Airport [1] are also provided.

Test Set	mAP	Δ mAP	Rank-1	Δ Rank-1
ENTIRE-ID [5]	43.6	+4.7	43.1	+4.7
DukeMTMC [4]	54.3	+4.1	72.3	+3.4
CUHK03 [25]	42.7	+4.1	44.3	+3.9
Market-1501 [3]	87.3	-2.3	94.6	-0.9
Airport [1]	69.4	+49.4	63.1	+40.7

As can be seen in Table 4.5, the mAP score over all datasets improved by

between 4.1 and 4.7 and, similarly, the CMC Rank-1 improved by between 3.4 and 4.7 for the model trained without REA. Meanwhile, adding Airport [1] to the dataset did not considerably hinder performance on the original Market-1501 [3] dataset, which dropped by 2.3 (mAP) and 0.9 (Rank-1). Naturally, the performance on the Airport [1] dataset was greatly improved. It is worth noting that, in order to exploit the differences between Market-1501 [3] and Airport [1] as much as possible, we created the training set by adding every Airport image annotated with a valid and reappearing person ID to Market’s training set, so the scores obtained by the model on Airport [1] should not be given much weight, they were added for the sake of completeness. Finally, as it was expected, the cross-dataset scores receive a slight improvement when removing random erasing augmentation from the training process, except on the ENTIRE-ID [5] dataset, where the model trained with REA obtained, surprisingly enough, higher cross-dataset scores. This is possibly due to the fact that the people represented in the ENTIRE-ID dataset [5] have their faces blurred, as can be seen in Figure 1.4, which is a form of artificial occlusion not too dissimilar from the one produced by REA, and thus it is possible that the models trained using it were better equipped to handle the ENTIRE-ID dataset [5].

4.2.3 Additional Experiments

To further investigate the factors contributing to performance improvements, we extended our experiments by including additional datasets in the training process. This allowed us to assess whether the observed performance gains were driven solely by the increase in the quantity of training data, or if strategically selecting datasets based on cross-dataset scores to identify and include the most challenging or ”distant” datasets, as was done in Section 4.2.2, leads to more effective performance improvements compared to simply adding more data indiscriminately.

To this end, we created two different training datasets, starting from the union of Market-1501 [3] and Airport [1] described in Section 4.2.2:

- In the first union, we added the training set from DukeMTMC-reID [4].
- In the second union, we added the training set from CUHK03 [25] and the test set from ENTIRE-ID [5].

As was done in Section 4.2.2, we trained the ViT CLIP-ReID model [123], both with and without REA (Section 4.1.2) and repeated the cross-dataset tests for with the remaining datasets. The results are presented in Table 4.6 (for the model trained on the first union) and Table 4.7 (for the model trained on the second union).

Table 4.6: Cross-dataset test result for the ViT CLIP-ReID model [123] trained on the union of Market-1501 [3], Airport [1] and DukeMTMC-reID [4].

Rea?	ENTIRE-ID [5]		CUHK03 [25]	
	mAP	Rank-1	mAP	Rank-1
Yes	50.6	50.1	41.6	43.1
No	48.2	47.2	43.1	44.7

Table 4.7: Cross-dataset test result for the ViT CLIP-ReID model [123] trained on the union of Market-1501 [3], Airport [1], ENTIRE-ID [5] and CUHK03 [25].

Rea?	mAP	Rank-1
Yes	54.5	70.7
No	56.6	73.1

For both unions, the additional data seems to improve the cross-dataset scores, however the improvements are reduced when compared to the ones obtained by simply adding Airport (Table 4.5). In particular, adding DukeMTMC-reID [4] only improved the performance on CUHK03 [25] by 0.4% (mAP) and

0.4% (Rank-1). The additional data had the greatest impact on the cross-dataset scores of the ENTIRE-ID dataset [5], where once again REA seems to help the cross dataset scores rather than hinder them. When using REA, adding DukeMTMC-reID [4] to the training set improved the performance by 4.4% (mAP) and 4.2% (Rank-1); without REA, they improved by 4.6% (mAP) and 4.1% (Rank-1).

Similarly, adding CUHK03 [25] and the test set of ENTIRE-ID [5] only improved the performance on DukeMTMC [4] by 2.3% (mAP) and 0.8% (Rank-1).

Adding more datasets to the union consistently resulted in a smaller improvement in performance (the only close score improvement being the one obtained on ENTIRE-ID [5] without REA). These findings suggest that leveraging cross-dataset evaluations to strategically identify and incorporate the most challenging dataset is more effective for improving model performance than merely increasing the quantity of training data without considering dataset relevance.

Proposed Dataset Selection Strategy

Motivated by these observations, we addressed the problem of selecting the most suitable training datasets for a given target dataset D_t from a collection of homogeneous candidates by proposing a *Dataset Mining Method* that leverages performance on the target as the key selection criterion. The method proceeds as follows:

- Train a separate baseline model (with the same architecture) on each candidate source dataset.
- Evaluate each model on its own validation set and on a labeled subset of D_t to measure the performance gap. If labels for D_t are unavailable, the gap can be approximated by comparing embeddings generated on D_t , using similarity or distance measures to estimate domain proximity. Next, select the dataset and its related model with the *smallest*

performance gap between source validation and target.

- Evaluate the selected model on all other source datasets (cross-source evaluation) and identify the dataset where this model performs worst. By transitivity, this dataset is also expected to have a large gap from the target.

This approach balances stability and diversity. Experiments (Tables 4.4 and 4.2) show that combining Market-1501 (smallest gap from the target) and Airport (on which the Market-1501 trained model performs worst) gives the highest gain (+4.3% mAP), while intermediate-gap datasets contribute less.

This method is inspired by similar principles in related contexts. For instance, catastrophic forgetting and lifelong learning approaches, such as DJAA[173], select embeddings to preserve prior knowledge while encouraging adaptation: embeddings close to previously learned prototypes prevent forgetting, while harder examples guide continual learning. Similarly, Batch Hard Mining[16] focuses on the most challenging negatives within a batch to improve feature robustness. Our method adapts these principles to the problem of dataset selection: the small-gap dataset aligns features with the target, providing stability, while the large-gap dataset introduces challenging examples that drive generalization.

4.2.4 Conclusions

Despite the rapid advancements in the field of person re-id [9], [10], [14], [205], deploying these models in real-world settings with the same high accuracy seen in benchmark environments remains a challenging goal. The domain gap, or the performance drop when models are applied to data that differs significantly from their training datasets, continues to be a primary obstacle.

In this study, we experimentally measured the effects of the domain gap on state-of-the-art models [123] using various benchmarking datasets, including

Airport [1], Market-1501 [3], and DukeMTMC-reID [4]. Our results demonstrate that CLIP-ReID [123] models, while excelling in controlled benchmark scenarios, experience significant performance degradation when confronted with datasets that differ substantially from their training data. This finding underscores the limitations of current models in adapting to diverse, real-world environments.

To address this challenge, we experimented with re-training the best-performing model on a combined dataset of the two most distinct benchmarks in our study. This approach aimed to reduce the domain gap without requiring a dramatic increase in labeled data. Our re-training method achieved an average improvement of +4.3% in mean Average Precision (mAP) and +4.0% in Cumulative Matching Characteristic (CMC) Rank-1 accuracy, showing a modest yet meaningful enhancement in model robustness across different domains.

While these results represent a positive step, they highlight the ongoing need for innovative domain adaptation strategies to make person re-id models truly adaptable to varied real-world settings. Future work should continue exploring techniques that enable models to generalize across diverse environments, potentially through few-shot learning or incorporating more varied data sources. With recent advancements in generative AI, unsupervised domain adaptation—leveraging the generation of synthetic variants of specific identities—could enhance model generalization without extensive labeled data. Utilizing such generative approaches for domain translation [98] and feature alignment [215] is a promising path toward making re-id models more resilient to environmental variability, ultimately supporting more reliable deployment across a wide range of operational scenarios.

4.3 Conclusions on Research Questions

4.3.1 Insight on research question 3

RQ: How robust are current state-of-the-art person re-identification models to domain shifts, and what is the effectiveness of domain adaptation strategies in bridging performance gaps?

This chapter presents a proof-of-concept addressing this question through quantitative cross-dataset evaluations under various scenarios. We experimented with a CLIP-based framework fine-tuned for person re-identification [123] and assessed the impact of Random Erasing Augmentation techniques [19].

The results, summarized in Tables 4.2, 4.4, and 4.5, show that while Random Erasing improved robustness on some datasets, it led to performance degradation on others. Our analysis suggests that the effectiveness of this augmentation depends on specific dataset characteristics, such as the prevalence of blurred faces.

In conclusion, our findings [99] highlight the complexity of domain adaptation in person re-identification and provide a partial quantification of its effectiveness. The work highlights two main strategies to bridge domain gaps: using more challenging or synthetic data, and advancing training through cross-dataset evaluations. This suggests that addressing domain shifts in person re-identification is largely a data problem rather than a limitation of model capacity.

4.3.2 Insight on research question 4

RQ: What is the practical impact of person re-identification models in multi-pedestrian tracking systems, and how can their integration be optimized?

The study [99] provides partial insights into the practical impact of using person re-identification (Re-ID) models within multi-pedestrian tracking

systems [28], [52], which we will further examine in the next chapter.

Tracking systems typically employ Re-ID models at multiple stages in their decision-making pipeline. At each stage, the number of identities to discriminate can vary considerably ([216]) — ranging from only a few individuals in the frame to a large gallery of identities with multiple samples per identity in long-term tracking.

The experiments reported in [99] show a notable drop in performance when state-of-the-art Re-ID models are applied to data distributions different from their training sets ([123], Table 4.2). However, in short-term tracking scenarios, where the number of identities to match is small, even lower-performing models can provide meaningful cues for tracking.

This highlights that optimizing Re-ID models for the specific tracking context is crucial to maximize their practical utility in real-world benchmarks.

Chapter 5

At the intersection of Person ReID and Tracking

This chapter presents a critical and systematic investigation into the intersection between Person Re-Identification (ReID) models and Multi-Person Tracking (MOT) systems [216]. The research began during my tenure as a Visiting Researcher at INRIA, Sophia Antipolis, under the supervision of Professor François Brémond, and has continued independently since then. While certain components of this work remain subject to future investigation, the experiments conducted provide significant empirical insights.

Recent state-of-the-art tracking algorithms, including DeepOCSort [217], StrongSORT [218], Ghost [219], and BoTSort [28], increasingly rely on Person ReID features as auxiliary cues for data association, highlighting the relevance of this intersection.

Previous publications on both ReID [14], [95], [98] and tracking [28], [52], [217], [218] have highlighted a saturation in parameter tuning required to achieve top results on benchmarks such as [3], [216]. This phenomenon is typical when benchmarks have existed for over a decade in a highly researched field.

However, a systematic evaluation of model performance specifically at the intersection of ReID and tracking systems appears to be lacking.

This chapter, along with the planned future work, aims to contribute to this narrative by:

- Conducting a comprehensive set of experiments that explore interactions between ReID and tracking, ranging from simple image-based benchmarks to video-based models and integration into multi-person tracking pipelines.
- Providing a critical analysis that quantifies the generalization and robustness of current ReID models when embedded in multi-person tracking systems, highlighting factors that affect performance beyond static benchmarks.

The experimental workflow is organized into three main sections. Section 5.1 discusses image-based ReID evaluation. A ReID backbone [30] was implemented, trained, and tested on standard image-based ReID benchmarks (Market [3]) to establish a baseline of feature discriminative capability.

Section 5.2 presents video-based ReID evaluation. The backbone was extended to video scenarios, training and testing on video ReID datasets to capture temporal information. The primary results achieved state-of-the-art performance on the iLIDS-VID dataset [2].

Section 5.3 presents the integration and adaptation of the model into the BotSort [28] tracking pipeline. We then evaluate the joint tracking system using standard MOT metrics and protocols. Each section includes detailed descriptions of the experimental design, implementation, and results analysis.

Overall conclusions and insights are drawn in Section 5.4.

5.1 Image-based ReID evaluation

The recent work SEAS Zhu, Budhwant, Zheng, *et al.*[30], presented at CVPR 2024, achieved impressive results on standard ReID benchmarks. The paper reports 98% for both *mAP* and *Rank-1* on Market-1501 [3], and 92% *mAP* and 91% *Rank-1* on the MSMT17 dataset [26].

Since the original implementation of the paper was not publicly released at the time of this research, we developed an independent implementation following the architectural specifications described in SEAS [30]. The backbone follows the rules of [19], a backbone well known in literature of Person ReID as "Strong-Baseline architecture". We introduced already in Section 2.2.2 as Equation 2.1 and Equation 2.4. We reported its explanation in the caption of the Figure 5.1 below.

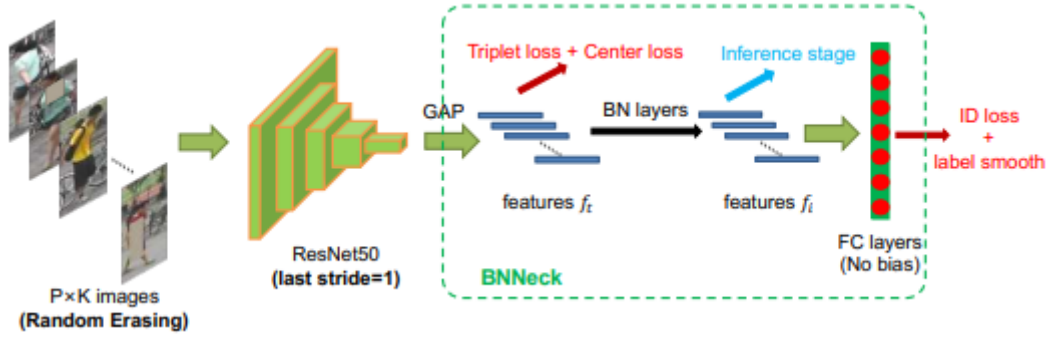


Figure 5.1: Image source [19]. A ResNet50[22] backbone pretrained on ImageNet [53] (and stride=1 in last layer) extracts features, reduced via Global Average Pooling (GAP). Embeddings are trained with Triplet and Center losses (Section 2.2.2 Equation 2.4) sampled with *HardBatchMining* strategy [16], plus ID loss (cross-entropy) with label smoothing (Section 2.2.2, Equation 2.1). Data augmentations (random flip, random erasing) improved training. At inference, embeddings after BatchNorm layers are used.

In SEAS ([30]), the *Strong baseline architecture*[19] was enhanced with an additional loss to guide the identity encoder in generating pixel-level body shape features. This was achieved by introducing a trainable body shape network that converts the identity feature maps of the baseline into a compressed 3D body shape representations. This loss is able to provide shape-aware supervision at the pixel level without increasing computational cost at inference time. For reference, focus on the orange rectangle on the left in Figure 5.2.

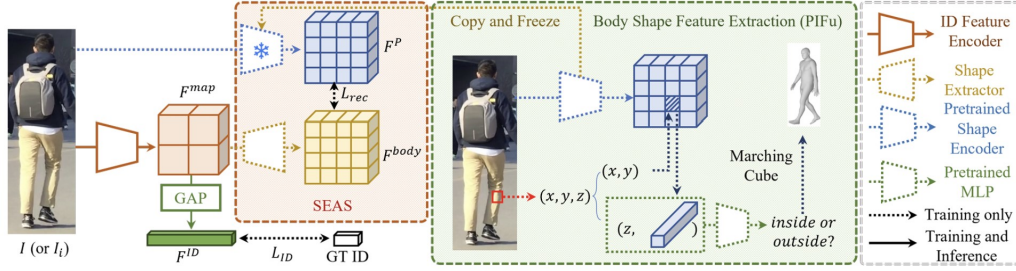


Figure 5.2: Image and text source [30]. On the left, the SEAS pipeline showing how the Person Re-Identification uses the body shape extractor output in supervision with the implicit shape encoder. The reconstruction loss is applied. On the right side of the image, a brief overview of the PIFu [220] pipeline. The z is a fixed value given with the data to represent the camera depth. Dotted lines means trainable and Trapezoids are neural networks. GAP stands for Global Average Pooling.

The added reconstruction loss is denoted in the figure as $\mathcal{L}_{rec}$ and its formula is:

$$\mathcal{L}_{rec} = \frac{1}{HW} \sum \text{SmoothL1}(F^{body}, F^P) \quad (5.1)$$

where H and W are respectively the height and width of the output tensor to compare: F^{body} which is the output of the trainable encoder, and F^P , the output of the frozen PIFu [220] backbone, discussed hereafter.

The total loss in SEAS pipeline for image-based reidentification is

$$\mathcal{L} = \mathcal{L}_{triplet} + \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{rec} \quad (5.2)$$

With λ_1 equal to 0.1.

On the right side (green rectangle) of Figure 5.2, an overview of the model whose inference output (F^P) is used as ground truth in the $\mathcal{L}_{rec}$ loss (Eq. 5.1) is shown.

This model was originally trained for the PIFu framework [220], computing feature vectors for each pixel, which are then combined with the camera depth (denoted as z in the image) and fed into a layer to predict the signed distance function (SDF). PIFu aims to reconstruct clothing surface geometry

by using the SDF, which assigns a scalar value to each pixel representing the shortest distance to the surface: positive values indicate points outside, and negative values indicate points inside the surface, enabling precise per-pixel depth estimation. In the original work, the resulting SDF is converted then into a 3D mesh using the Marching Cubes algorithm [221].

The equation below provides a formal mathematical description of the process illustrated in the figure:

$$\text{PIFU}(F_i^P, z) = s, \quad s \in \mathbb{R}, \quad (5.3)$$

where F_i^P is the feature vector of pixel i , and z represents the camera depth, which is fixed during training.

In SEAS [30] Zhu, Budhwant, Zheng, *et al.* argued that the output of the PIFU encoder [220] captures, at each pixel, salient information for those scalars s and, on the clothing surfaces as well. For this reason, it can be effectively used as ground truth for the loss, ensuring the model focuses on these human shape cues.

In order to evaluate the pure capability of the model to discriminate among images, we tested it on Market1501[3] and reported the results in Table 5.1, also, we added the original paper declared results and the [222] FastReID values for consistency with further experiments on tracking: Zheng, Shen, Tian, *et al.*, [3].

Market1501	mAP	Rank-1
Our	82.2	93.9
Our w/ Reranking	92.5	95.2
[30] Seas	98.9	98.6
[222] FastReID	88.7	95.2

Table 5.1: Models evaluation on Market1501[3].

Although our implementation achieves state-of-the-art results in *mAP* when the technique of re-ranking [223] is applied, as we will discuss later,

the model still performs well in contexts where only the first match count, achieving 93.9% in *Rank-1 accuracy*

5.2 Video-based reid evaluation

In this section we evaluated the ability of the model to discriminate with temporal aggregations. For Video-based Re-Identification we evaluated two different solutions: the PSTA [32] and the aggregation by Transformers. Both works aggregating the embedding of each frame in a sampled sequence. To a complete overview on the differences between image and video sequences, strengths and weaknesses, we remand to Section 1.1.2 of this document.

Technical Details on PSTA

Pyramid Spatial-Temporal Aggregation (PSTA) [32], developed by Wang, Zhang, Gao, *et al.*, was presented at ICCV 2021. The technique aggregates information from both space (within a frame) and time (across frames) using a pyramid structure and attention mechanisms. The overview is reported in Figure 5.3.

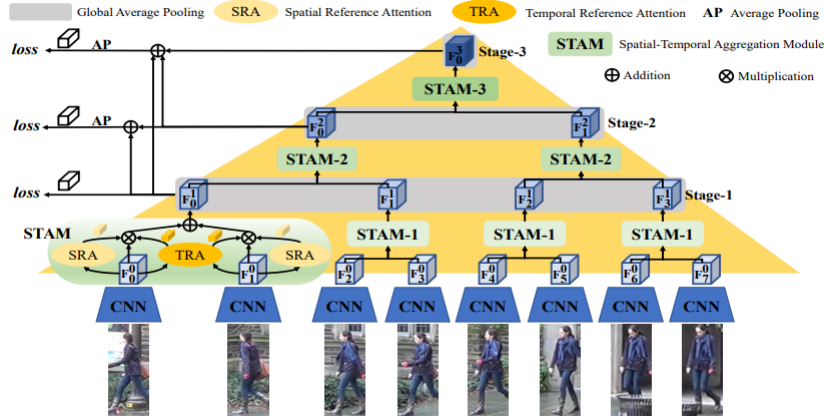


Figure 5.3: Image source [32]. Overall structure of our PSTA framework with 8 frames as example. STAM shares both structure and parameters within each stage of the pipeline.

In the framework, a sequence of eight frames in a tracklet are sampled randomly and uniformly. As showed in the image 5.3, once the features are extracted by a generic backbone, the pipeline progressively aggregates frame-level features in a hierarchical-pyramidal way: it first combines features from nearby frames (short-term), then fuses these into longer-term representations, allowing the model to capture both short-term and long-term temporal dependencies.

The algorithm work with three modules, namely Spatial-Temporal Aggregation Module (STAM), Spatial Reference Attention (SRA) and Temporal Reference Attention (TRA):

- STAM: Aggregates features by sequentially applying the TRA and SRA modules, and on their output it applies convolutional and batch normalization layers to refine the representations.
- SRA: Computes channel-wise spatial attention maps for each frame to highlight important regions, multiplies these weights by the original features, and fuses each pair of adjacent frames using non-linear activation, resulting in an output sequence half as long as the input.
- TRA: For every pair of adjacent frames, computes attention maps that highlight consistent and important information across time, then fuses the two frames' features using these learned weights, producing a temporally-aggregated feature robust to noise and occlusion.

The aggregation process is repeated three times since every sequence has eight frames, resulting in a three-level pyramidal aggregation.

Technical Details on Transformer

The use of transformers for embedding aggregation in video-based person re-identification has become an active research area. The core architectural pattern—using a learnable CLS token to aggregate temporal information—originates from BERT [224] in NLP and ViT [23] in computer vision.

Early applications to video-based re-ID include TCLNet [225] and CAViT [226] (both 2021-2022), which specifically used transformer encoders with CLS tokens for temporal aggregation across video frames. The idea of omitting strict positional encoding to enable permutation-invariant frame aggregation, treating videos as sets rather than ordered sequences—an approach conceptually aligned with Set Transformer [227].

Some recent work use the transformers in the field, but none of them used them in this way at the best of our knowledge at the writing time of this thesis. Recent advances include PeT-KeyStAtion (Kim, Cho, Lee, *et al.*, 2025) [228], which uses keypoint-guided spatial-temporal aggregation to focus on salient regions, and Spatio-temporal Feature-level Augmentation ViT (Ma, Yi, and Gall, 2025)[229], which applies augmentation directly to feature embeddings for improved robustness.

Our transformer-based aggregator model follows this established pattern with specific design choices:

Architecture. The model is an encoder-only transformer that processes sequences of $T = 8$ frame embeddings. Each input embedding $\mathbf{x}_t \in \mathbb{R}^{2048}$ is first projected to model dimension via a linear layer: $\mathbf{h}_t = \mathbf{W}_{\text{proj}}\mathbf{x}_t$, where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{2048 \times 1024}$. A learnable CLS token $\mathbf{c} \in \mathbb{R}^{1024}$ is prepended to form the sequence $[\mathbf{c}; \mathbf{h}_1, \dots, \mathbf{h}_T]$.

Transformer Encoder. The sequence undergoes layer normalization followed by processing through 4 transformer encoder layers. Each layer uses 16 attention heads ($d_{\text{head}} = 64$) and a feedforward network with hidden dimension 2048, configured with pre-normalization (`norm_first=True`) and without bias terms. Critically, *no positional encoding* is applied, making the model permutation-invariant to frame order. This design choice treats video frames as an unordered set and provides robustness to random temporal sampling during training.

Output Processing. The final CLS token representation $\mathbf{c}_{\text{out}} \in \mathbb{R}^{1024}$ is projected back to 2048 dimensions: $\mathbf{f}_{\text{global}} = \mathbf{W}_{\text{out}}\mathbf{c}_{\text{out}}$. Following the "Bag of Tricks" framework [19], the global feature undergoes batch normalization

(with bias frozen to zero) to produce the normalized feature $\mathbf{f}_{\text{norm}}$, which is then used for both metric learning and classification via a linear classifier $\mathbf{W}_{\text{cls}} \in \mathbb{R}^{N \times 2048}$ (without bias), where N is the number of training identities.

Evaluation

We evaluated the two techniques on the previous built image based Person ReID model. We reported the results on Table 5.2 along with the values of the original cited papers [30], [32].

The results demonstrate that our transformer implementation achieves state-of-the-art performance on iLIDS-Vid, obtaining a mAP of 88.2% and Rank-5 accuracy of 96.7%, surpassing existing methods including PSTA[32] (85.8% mAP) and SEAS [30] (79% mAP) as they reported from their papers.

Test on iLIDS-Vid	mAP	Rank-1	Rank-5
[30] Seas (from paper)	79	83,5	94
[32] PSTA (from paper)	85,8	91,5	-
myPSTA impl.	80,3	73,3	88,7
*myTrans. impl.	88.2 (σ :0.9)	81.7 (σ :1.5)	96.7 (σ :0.8)

Table 5.2: Models evaluation on test set of iLIDS-Vid[2]. σ : avg std on 5 runs.

The most significant finding of this work is that even though the image-based backbone is not the best, our video-based method outperforms others by effectively leveraging temporal information. The transformer combines spatial and temporal cues in videos so efficiently that it compensates for the backbone’s limitations. Moreover, the low standard deviation values indicates that the method is robust and stable in its results.

The success of video-based methods with weaker backbones shows that temporal dynamics play a critical role, alongside spatial features, for optimal re-identification. Future research should focus on balancing spatial and tem-

poral modeling, testing generalizability, and developing efficient video-based architectures.

5.3 MOT Tracking Benchmark evaluation

As introduced, the Person ReID feature as auxiliary cues for data association, is becoming an adopted paradigm among modern tracking algorithms [28], [217]–[219]. In MOT17 challenge [216] benchmarks, Aharon, Orfaig, and Bobrovsky stands out with the BoT-SORT algorithm [28]. By integrating our re-ID model into BoT-SORT, we aim to investigate whether our reid-model can affecting the tracking accuracy and identity consistency of the algorithm.

Technical Details

The BoT-SORT[28] pipeline is presented in Figure 5.4. Each detection in a new frame (Yolo[51]) is first matched to existing active tracks using motion (Kalman filter[230]) and appearance (Re-ID [222]) cues . If a detection cannot be associated with any active track, the algorithm attempts to match it with tracks that are currently lost (i.e., tracks that have not been updated for a few frames but are still retained in memory). When a lost track is successfully matched to a new detection, it is immediately reactivated as an active track. If a detection does not match any active or lost track, a new track is created in an unconfirmed state; this track becomes confirmed only after being matched with detections in consecutive frames. Tracks that remain unmatched for a predefined number of frames are removed from memory, ensuring that only relevant and recently observed objects are tracked.

Evaluation

For our experiments on MOT17 Challenge[216], we used TrackEval¹[231], a Python library designed from the benchmark to evaluate multi-object track-

¹<https://github.com/JonathonLuiten/TrackEval>

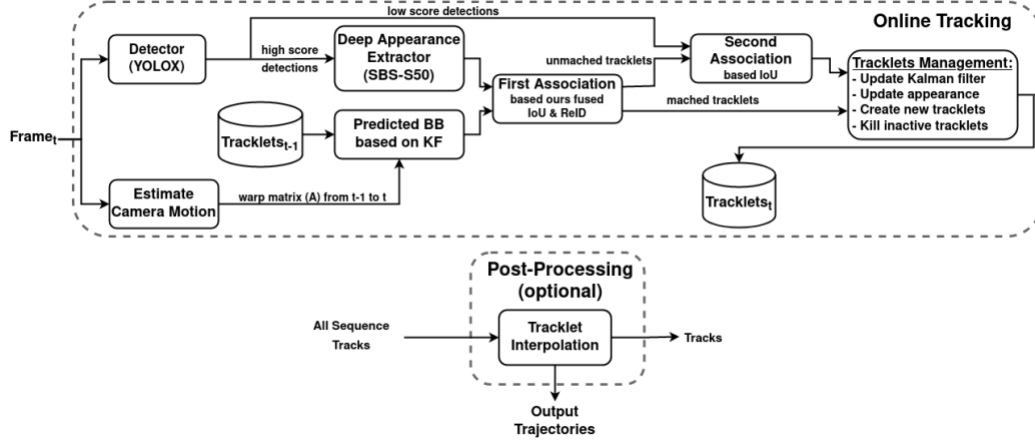


Figure 5.4: Image source [28]. The pipeline of BoT-SORT Algorithm is: detection → camera motion compensation → motion prediction → feature extraction ReID → association → sequences update.

ing (MOT) algorithms. Among the metrics, MOTA, HOTA and IDF1 are the most important:

- MOTA (Multi-Object Tracking Accuracy): measures how well the system tracks all objects in the video, accounting for missed targets, false positives, and identity switches.
- IDF1 (ID F1-Score): measures how well the system keeps the same identity for each object throughout the video. A higher score means fewer identity switches.
- HOTA (Higher Order Tracking Accuracy): evaluates both how accurately objects are detected and how consistently their identities are maintained over time.

MOT17 Fine Tuning Strategy: For our experiments, we refer to our implementation of the model trained on Market[3] using the SEAS[30] strategy as mySeas. This model was fine-tuned on the MOT17 dataset[216] following the temporal splitting strategy proposed by BoT-SORT[28] for training the FastReID[222] baseline.

Specifically, for each sequence in MOT17, the first half of frames was used for training and the second half for validation. Person IDs were kept global across all sequences by incrementing an *id_offset* equal to the *max_id_per_seq* after each sequence.

The model pre-trained on Market was fine-tuned using only the Triplets and Identity loss of *Strong baseline*[19], and an additional update to the batch normalization layer weights was applied before starting training, as reported in [219]. It is worth noting that FastReID[222] reported Rank-1 accuracy on Market very similar to the values in Table 5.1.

MOT17 Validation	HOTA	IDF1	MOTA
StrongSORT baseline	69.6	82.3	77.1
DeepOCSort baseline	<u>70.20</u>	82.78	-
BotSort w/ fastreid t.on MOT17 + interp. (Baseline)	69.17	82.07	78.46
BotSort w/ mySeas t.on MOT17 + interp. (ours)	70.295	<u>82.513</u>	80.455
BotSort w/ mySeas t.on MOT17	69.09	81.47	<u>78.47</u>
BotSort w/ mySeas	69.03	81.39	78.44

Table 5.3: Ablation study on the MOT17 validation set. Best results are in bold; second-best results are underlined; baseline and ours implementation are highlighted for making them easily readable and comparable.

Notably, the integration of our ReID model (mySeas) into BoTSort, with interpolation, improves all major metrics (HOTA: 70.295, MOTA: 80.455), surpassing both the StrongSORT and DeepOCSort baselines as well as the BoTSort version with FastReID. The ablation study reported in the table confirms also that the combination of the ReID model fine tuning and temporal interpolation is key to this performance leap.

On the test set, BotSort with our pipeline maintains competitive performance: it achieves the highest MOTA (80.65) and delivers results aligned with other methods on HOTA and IDF1. This demonstrates that the system is robust and generalizes well to unseen data, maintaining a balance between

MOT17 Test set	HOTA	IDF1	MOTA
StrongSORT	64.4	79.5	79.6
DeepOCSort	64.9	80.6	79.4
BotSort (baseline)	65	80.2	80.5
BotSort (ours)	64.65	79.5	80.65

Table 5.4: Evaluation on the MOT17 testing set. Best results are in bold. HOTA focus on both detection and identities accuracy. MOTA focus on detection accuracy, IDF1 focus on preserving identities

detection accuracy and identity preservation despite the model being 10% less performant on mAP for the static content.

We analyzed specific weaknesses in the three worst-performing MOT17 test sequences. In *MOT17-14* there is a busy intersection of people recorded by a camera on a moving bus. Low association (43.5%) and MOTP (43.5%) indicate that the tracker fails under fast motion and Re-ID on hard boxes. Sequence *MOT17-08* shows a crowded pedestrian street that is recorded by a stationary camera: the Re-ID features are corrupted by occlusions and overlapping boxes, leading to very high identity switches (1086) and low detection recall (47.3%). Sequence *MOT17-01* presents people walking around a large square: detection performance was moderate (recall 54.5%), but identity association remained challenging (IDF1 61.2, AssA 52.9) due to freely moving pedestrians. These observations align with Stanczyk and Bremond (2025)[232]: in "in-the-wild" scenarios, Re-ID is beneficial only when contextual constraints are applied, such as occlusion-aware attention or bounding box filtering. Without these, Re-ID risks introducing noise rather than resolving ambiguities.

Future research should move toward context-aware, adaptive fusion of Re-ID cues within MOT, where the contribution of appearance is dynamically weighted according to detection quality, motion reliability, and crowd density.

5.4 Conclusions and Discussions

In this chapter, we presented several insights from our research. We implemented and evaluated a discriminative Person Re-Identification model that achieved competitive *Rank-1 accuracy* in image-based ReID and demonstrated strong performance in video-based ReID and tracking scenarios.

For video-based ReID, we tested the image-based model using both PSTA and our Transformer-based approach, with the latter achieving state-of-the-art results. These outcomes suggest that our Transformer model is a strong candidate for future state-of-the-art performance, if confirmed by further experiments.

Our second line of investigation focused on instance discrimination within dynamic Multi-Person Tracking (MOT) systems. Here, the model outperformed the baseline on validation, with only minor drops in the three main MOT evaluation metrics during testing.

By examining the intersection between ReID and tracking, this work provides an ablation study that can serve as a methodological reference for future research. The findings emphasize that strong performance on isolated ReID or tracking tasks does not automatically translate to robustness when these components are integrated. Joint evaluation reveals important trade-offs between discrimination capability and temporal consistency.

5.5 Conclusions on Research Questions

As main insight we can see that a balanced equilibrium is required between spatial and temporal feature extraction, and only if we joint them with a robust domain adaptation we can maximize the practical impact in person re-identification models. Our key takeaway is that performance should not be assessed solely by benchmarks — a context-aware, data-driven perspective can turn a seemingly weak model into an effective and powerful solution.

5.5.1 Insight on research question 4

RQ: What is the practical impact of person re-identification models in multi-pedestrian tracking systems, and how can their integration be optimized?

The impact of adapted ReID models into multi-pedestrian tracking systems like BoTSort[28] provides measurable improvements in tracking metrics such as HOTA and MOTA[231], especially when temporal interpolation (used in validation) and domain-specific fine-tuning are applied.

The chapter’s ablation studies show that using a ReID model without adaptation results in suboptimal identity preservation and increased ID switches, while domain-adapted models significantly enhance the system’s ability to maintain consistent identities across frames.

However, the results also highlight that even strong ReID models can fail in highly dynamic or crowded scenes if not properly integrated, underscoring the need for joint optimization of appearance and motion cues. Thus, the chapter provides proof-of-concept that careful adaptation and integration of ReID models are critical for maximizing their practical impact in real-world multi-object tracking applications.

Chapter 6

Conclusions

6.1 Scientific Conclusions

This thesis provided a systematic, multi-perspective investigation of Person Re-Identification (ReID), addressing key challenges through theoretical, empirical, and application-level analysis. Four research axes guided the study: *RQ1* identified challenges and trends in supervised and unsupervised ReID; *RQ2* explored embeddings emphasizing identity-specific features while minimizing transient attributes; *RQ3* assessed robustness to domain shifts and adaptation strategies; *RQ4* evaluated integration in multi-pedestrian tracking systems.

The work aims to bridge the gap between core ReID methodologies and real-world multi-person tracking scenarios. The results provide empirical evidence on both the limitations and potential of current approaches, and highlight avenues for future research, offering contributions to the scientific discourse, technological development, and ethical considerations of the field.

The literature review (Chapter 2) showed that supervised ReID methods have reached near-saturation performance (mAP $\approx 98\%$ on Market1501[3] and DukeMTMC[4]), driven by effective loss functions on ResNet[22] and ViT[23] based architectures. Remaining challenges include robustness to domain shifts, noise in unsupervised pseudo-labels, and extraction of stable

identity-specific features. Fully unsupervised methods show a gap of 6% on Market1501[3] and 13% on DukeMTMC[4]. Promising directions include teacher-student, momentum-based, and self-paced learning strategies, as well as pre-trained models like CLIP[123].

In Chapter 3, the proof-of-concept on Denoising Diffusion Models [98] demonstrated that identity can be encoded as conditional information to generate images of the same person, separating it from pose and background. Using the right inverse networks of the generator for latent identity and context extraction, the model achieved 73% mAP on Market1501[3]—below supervised state-of-the-art but promising because of the low 64×32 resolution. Limitations include reliance on supervised training, while unsupervised extensions remain open. Results confirm the approach separates stable identity features from transient noise, showing the potential of modeling and removing transient components for identity-consistent representations.

The second proof-of-concept in Chapter 4 systematically assessed domain gaps [99], showing that even high-performing models drop significantly on unseen datasets. Re-training on strategically combined datasets improved mAP by +4.3% and Rank-1 accuracy by +4.0%, indicating that careful dataset selection outperforms mere data volume increase. CLIP-ReID[123] models, despite multimodal pre-training, remain vulnerable to domain shifts. These results highlight that domain adaptation is primarily data-centric, suggesting generative AI could create synthetic domain-specific variants without extensive manual annotation.

In Chapter 5, integrating ReID models into Multi-Person Tracking provided insights into practical applicability. Our video-based Transformer achieved state-of-the-art on iLIDS-VID [2](88.2% mAP), surpassing PSTA [32](85.8%) and SEAS [30](79%), showing temporal aggregation compensates for image-based backbone limits. Optimized integration in BoTSort[28] improved MOT17 metrics (HOTA: 70.295, MOTA: 80.455), outperforming StrongSORT[218] and DeepOCSort[217], with competitive test performance (MOTA: 80.65). Ablations confirm domain adaptation and temporal interpo-

lation are key. Results highlight that high task-specific performance does not guarantee robustness, stressing joint optimization of appearance and motion cues. However, from the experiments is evident that high reid performance on isolated tasks does not automatically translate into robustness when components are integrated, there is the need for joint optimization of appearance and contextual cues.

This thesis demonstrates that successful applied Person Re-Identification requires balancing three dimensions: discriminative feature capacity, robustness to domain shifts via strategic adaptation, and optimized integration into end-to-end systems like multi-object tracking. Performance should not rely on isolated benchmarks rather a context-aware, data-driven approach that can make models effective in real-world scenarios. Promising future directions include robust unsupervised methods, generative models for data augmentation and domain adaptation, multimodal pre-training with temporal aggregation, and continual domain adaptation. While the present work focuses on methodological and empirical contributions, a few thoughts on the ethical implications of ReID technologies are discussed in the next section.

6.2 Ethical e Social Considerations

While surveillance technologies existed prior to AI, deploying Person ReID models at scale introduces unprecedented ethical and societal implications for individual privacy.

It was 2016 when the Center on Privacy & Technology at Georgetown Law School presented a study on perpetuallineup.org¹, reporting that one in two American adults was included in a law enforcement face recognition network.

Having access to the data in governments records, matching faces from video streams is increasingly feasible. Under which conditions this operation is lawful and ethically justified is unclear or partially explained from the

¹available on October 23,2025

Europe Union by two regulations: the European Union’s General Data Protection Regulation (GDPR) [233] and the AI Act (2021) [234]. The GDPR emphasizes consent, data minimization, and purpose limitation, while the AI Act provides harmonized guidelines for trustworthy AI systems [235].

From a technical perspective two techniques that have significant impact on ethical and legal implications involve: the *embedding encryption and accessibility* and the *data processing location*. Embeddings are compact representations of some input data, so a possible approach is to treat them in the same way as the original input under GDPR regulations. So, as a consequence, encrypt their stored forms when necessary or process the data in specific location to supports privacy-preserving operation.

For instance, a use case that present both is for the case of [7] where transit stations can collect Origin-Destination data to optimize public transportation without identifying individuals. The plain images cannot be stored at all. the embeddings should be created near the camera networks and then encryption could ensure lawfulness. in this kind of software architectures.

The goal as humans and computer scientist, is not only regulatory compliance but also engineering AI that aligns innovation with human rights, transparency, and accountability [236]. Achieving this requires collaboration between researchers, regulators, and industry stakeholders to ensure that ReID serves society responsibly—without compromising privacy, dignity, or trust.

Bibliography

- [1] M. Gou, Z. Wu, A. Rates-Borras, O. Camps, R. J. Radke, *et al.*, “A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 3, pp. 523–536, 2018.
- [2] M. Li, X. Zhu, and S. Gong, “Unsupervised tracklet person re-identification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 7, pp. 1770–1782, 2020. **doi:** 10.1109/TPAMI.2019.2903058.
- [3] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1116–1124. **doi:** 10.1109/ICCV.2015.133.
- [4] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in *Computer Vision – ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds., Cham: Springer International Publishing, 2016, pp. 17–35. **doi:** 10.1007/978-3-319-48881-3_2.
- [5] S. Yıldız and A. N. Kasım, “Entire-id: An extensive and diverse dataset for person re-identification,” in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, 2024, pp. 1–5. **doi:** 10.1109/FG59268.2024.10581945.

- [6] A. M. Mansourian, V. Somers, C. De Vleeschouwer, and S. Kasaei, “Multi-task learning for joint re-identification, team affiliation, and role classification for sports visual tracking,” in *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, 2023, pp. 103–112. **doi:** 10.1145/3606038.3616172.
- [7] K. Huang, T. Azfar, J. Reilly, and R. Ke, “Transitoid: Transit OD data collection with occlusion-resistant dynamic passenger re-identification,” *CoRR*, vol. abs/2504.11500, 2025. **doi:** 10.48550/ARXIV.2504.11500. arXiv: 2504.11500. **url:** <https://doi.org/10.48550/arXiv.2504.11500>.
- [8] S. Gong, M. Cristani, S. Yan, and C. C. Loy, Eds., *Person Re-Identification* (Advances in Computer Vision and Pattern Recognition). Springer, 2014. **doi:** 10.1007/978-1-4471-6296-4.
- [9] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, “Deep learning for person re-identification: A survey and outlook,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, 2022. **doi:** 10.1109/TPAMI.2021.3054775.
- [10] W. Wei, W. Yang, E. Zuo, Y. Qian, and L. Wang, “Person re-identification based on deep learning — an overview,” *Journal of Visual Communication and Image Representation*, vol. 82, p. 103418, 2022. **doi:** <https://doi.org/10.1016/j.jvcir.2021.103418>. **url:** <https://www.sciencedirect.com/science/article/pii/S1047320321002765>.
- [11] N. K. Singh, M. Khare, and H. B. Jethva, “A comprehensive survey on person re-identification approaches: Various aspects,” *Multimedia Tools and Applications*, vol. 81, no. 11, pp. 15 747–15 791, 2022. **doi:** 10.1007/s11042-022-12585-w. **url:** <https://doi.org/10.1007/s11042-022-12585-w>.
- [12] X. Qian, W. Wang, L. Zhang, *et al.*, “Long-term cloth-changing person re-identification,” in *Computer Vision - ACCV 2020 - 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 - De-*

- cember 4, 2020, Revised Selected Papers, Part III*, H. Ishikawa, C. Liu, T. Pajdla, and J. Shi, Eds., ser. Lecture Notes in Computer Science, vol. 12624, Springer, 2020, pp. 71–88. **doi:** 10.1007/978-3-030-69535-4\5. **url:** https://doi.org/10.1007/978-3-030-69535-4%5C_5.
- [13] Y. Huang, Q. Wu, J. Xu, Y. Zhong, and Z. Zhang, “Clothing status awareness for long-term person re-identification,” in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, IEEE, 2021, pp. 11 875–11 884. **doi:** 10.1109/ICCV48922.2021.01168. **url:** <https://doi.org/10.1109/ICCV48922.2021.01168>.
- [14] A. Asperti, S. Fiorilla, S. Nardi, and L. Orsini, “A review of recent techniques for person re-identification,” *Machine Vision and Applications*, vol. 36, no. 1, p. 25, 2024. **doi:** 10.1007/s00138-024-01622-3. **url:** <https://doi.org/10.1007/s00138-024-01622-3>.
- [15] S. Horiguchi, D. Ikami, and K. Aizawa, “Significance of softmax-based features in comparison to distance metric learning-based features,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1279–1285, 2020. **doi:** 10.1109/TPAMI.2019.2911075. **url:** <https://doi.org/10.1109/TPAMI.2019.2911075>.
- [16] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” *arXiv preprint arXiv:1703.07737*, 2017.
- [17] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, IEEE Computer Society, 2015, pp. 815–823. **doi:** 10.1109/CVPR.2015.7298682. **url:** <https://doi.org/10.1109/CVPR.2015.7298682>.

- [18] Y. Li, R. Xue, M. Zhu, J. Xu, and Z. Xu, “Angular triplet loss-based camera network for reid,” in *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, IEEE, 2021, pp. 1–7. **doi:** 10.1109/IJCNN52387.2021.9534270. **url:** <https://doi.org/10.1109/IJCNN52387.2021.9534270>.
- [19] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, “Bag of tricks and a strong baseline for deep person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, Computer Vision Foundation / IEEE, 2019, pp. 1487–1495. **doi:** 10.1109/CVPRW.2019.00190. **url:** http://openaccess.thecvf.com/content%5C_CVPRW%5C_2019/html/TRMTMCT/Luo%5C_Bag%5C_of%5C_Tricks%5C_and%5C_a%5C_Strong%5C_Baseline%5C_for%5C_Deep%5C_Person%5C_CVPRW%5C_2019%5C_paper.html.
- [20] R. Vezzani, D. Baltieri, and R. Cucchiara, “People reidentification in surveillance and forensics: A survey,” *ACM Comput. Surv.*, vol. 46, no. 2, 2013. **doi:** 10.1145/2543581.2543596. **url:** <https://doi.org/10.1145/2543581.2543596>.
- [21] P. K. Sarker, Q. Zhao, and M. K. Uddin, “Transformer-based person re-identification: A comprehensive review,” *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 7, pp. 5222–5239, 2024. **doi:** 10.1109/TIV.2024.3350669.
- [22] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. **doi:** 10.1109/CVPR.2016.90.
- [23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

- [24] H. Luo, P. Wang, Y. Xu, *et al.*, “Self-supervised pre-training for transformer-based person re-identification,” *ArXiv*, vol. abs/2111.12084, 2021. **url:** <https://api.semanticscholar.org/CorpusID:244488587>.
- [25] W. Li, R. Zhao, T. Xiao, and X. Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159. **doi:** 10.1109/CVPR.2014.27.
- [26] L. Wei, S. Zhang, W. Gao, and Q. Tian, “Person transfer gan to bridge domain gap for person re-identification,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 79–88. **doi:** 10.1109/CVPR.2018.00016.
- [27] D. Fu, D. Chen, J. Bao, *et al.*, “Unsupervised pre-training for person re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 750–14 759. **doi:** 10.1109/CVPR46437.2021.01451. **url:** <https://doi.org/10.1109/CVPR46437.2021.01451>.
- [28] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, *Bot-sort: Robust associations multi-pedestrian tracking*, 2022. **doi:** 10.48550/arXiv.2206.14651. **url:** <https://arxiv.org/abs/2206.14651>.
- [29] T. He, X. Jin, X. Shen, J. Huang, Z. Chen, and X. Hua, “Dense interaction learning for video-based person re-identification,” in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 2021, pp. 1470–1481. **doi:** 10.1109/ICCV48922.2021.00152.
- [30] H. Zhu, P. Budhwant, Z. Zheng, and R. Nevatia, “Seas: Shape-aligned supervision for person re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 164–174. **doi:** 10.1109/CVPR52733.2024.00024. **url:** <https://doi.org/10.1109/CVPR52733.2024.00024>.

- [31] C. Liu, J. Chen, C. Chen, and S. Chien, “Video-based person re-identification without bells and whistles,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, Computer Vision Foundation / IEEE, 2021, pp. 1491–1500. **doi:** 10.1109/CVPRW53098.2021.00165. **url:** https://openaccess.thecvf.com/content/CVPR2021W/AMFG/html/Liu%5C_Video-Based%5C_Person%5C_Re-Identification%5C_Without%5C_Bells%5C_and%5C_Whistles%5C_CVPRW%5C_2021%5C_paper.html.
- [32] Y. Wang, P. Zhang, S. Gao, X. Geng, H. Lu, and D. Wang, “Pyramid spatial-temporal aggregation for video-based person re-identification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12026–12035. **doi:** 10.1109/ICCV48922.2021.01181. **url:** <https://doi.org/10.1109/ICCV48922.2021.01181>.
- [33] L. Zheng, Z. Bie, Y. Sun, *et al.*, “Mars: A video benchmark for large-scale person re-identification,” in *Computer Vision – ECCV 2016*, Cham: Springer International Publishing, 2016, pp. 868–884. **doi:** 10.1007/978-3-319-46466-4_52. **url:** https://doi.org/10.1007/978-3-319-46466-4_52.
- [34] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, “Object detection in 20 years: A survey,” *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023. **doi:** 10.1109/JPROC.2023.3238524.
- [35] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, “Towards real-time multi-object tracking,” in *Computer Vision – ECCV 2020*, Cham: Springer International Publishing, 2020, pp. 107–122. **doi:** 10.1007/978-3-030-58621-8_7. **url:** <https://github.com/Zhongdao/Towards-Realtime-MOT>.
- [36] W. R. Schwartz and L. S. Davis, “Learning discriminative appearance-based models using partial least squares,” in *2009 XXII Brazilian*

- Symposium on Computer Graphics and Image Processing*, 2009, pp. 322–329. doi: 10.1109/SIBGRAPI.2009.42.
- [37] W.-S. Zheng, S. Gong, and T. Xiang, “Associating groups of people,” in *British Machine Vision Conference (BMVC) 2009*, 2009, pp. 1–11. doi: 10.5244/C.23.23. url: <https://bmva-archive.org.uk/bmvc/2009/Papers/Paper167/Paper167.pdf>.
- [38] C. C. Loy, T. Xiang, and S. Gong, “Multi-camera activity correlation analysis,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1988–1995. doi: 10.1109/CVPR.2009.5206827.
- [39] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, “Custom pictorial structures for re-identification,” in *British Machine Vision Conference (BMVC) 2011*, 2011, pp. 1–11. doi: 10.5244/C.25.68. url: <https://bmva-archive.org.uk/bmvc/2011/Papers/CustomPictorialStructures.pdf>.
- [40] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, “Relaxed pairwise learned metric for person re-identification,” in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 780–793. doi: 10.1007/978-3-642-33783-3_56. url: https://doi.org/10.1007/978-3-642-33783-3_56.
- [41] D. Gray and H. Tao, “Viewpoint invariant pedestrian recognition with an ensemble of localized features,” in *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I*, D. A. Forsyth, P. H. S. Torr, and A. Zisserman, Eds., ser. Lecture Notes in Computer Science, vol. 5302, Springer, 2008, pp. 262–275. doi: 10.1007/978-3-540-88682-2_21. url: https://doi.org/10.1007/978-3-540-88682-2_21.

- [42] W. Li and X. Wang, “Locally aligned feature transforms across views,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3594–3601. **doi:** 10.1109/CVPR.2013.461.
- [43] L. Wu, Y. Wang, J. Gao, M. Wang, Z.-J. Zha, and D. Tao, “Deep coattention-based comparator for relative representation learning in person re-identification,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 722–735, 2021. **doi:** 10.1109/TNNLS.2020.2979190.
- [44] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017. **doi:** 10.1109/TPAMI.2016.2577031.
- [45] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010. **doi:** 10.1109/TPAMI.2009.167.
- [46] O. Camps, M. Gou, T. Hebble, *et al.*, “From the lab to the real world: Re-identification in an airport camera network,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 540–553, 2017. **doi:** 10.1109/TCSVT.2016.2556538.
- [47] P. Dollár, R. Appel, S. Belongie, and P. Perona, “Fast feature pyramids for object detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014. **doi:** 10.1109/TPAMI.2014.2300479.
- [48] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI’81, Vancouver, BC, Canada: Morgan Kaufmann Publishers Inc., 1981, pp. 674–679.

- [49] E. Rosten, R. Porter, and T. Drummond, “Faster and better: A machine learning approach to corner detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 105–119, 2010. **doi:** 10.1109/TPAMI.2008.275.
- [50] X. Shu, X. Wang, X. Zang, *et al.*, “Large-scale spatio-temporal person re-identification: Algorithms and benchmark,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4390–4403, 2022. **doi:** 10.1109/TCSVT.2021.3128214.
- [51] G. Jocher, A. Chaurasia, and J. Qiu, *Ultralytics YOLO*, version 8.0.0, 2023. **url:** <https://github.com/ultralytics/ultralytics>.
- [52] Y. Zhang, P. Sun, Y. Jiang, *et al.*, “Bytetrack: Multi-object tracking by associating every detection box,” in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham: Springer Nature Switzerland, 2022, pp. 1–21. **doi:** 10.1007/978-3-031-20047-2_1. **url:** https://doi.org/10.1007/978-3-031-20047-2_1.
- [53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. **doi:** 10.1109/CVPR.2009.5206848.
- [54] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior, “The relation between the roc curve and the cmc,” in *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID’05)*, 2005, pp. 15–20. **doi:** 10.1109/AUTOID.2005.48.
- [55] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, “Learning to rank in person re-identification with metric ensembles,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, IEEE Computer Society, 2015, pp. 1846–1855. **doi:** 10.1109/CVPR.2015.7298794. **url:** <https://doi.org/10.1109/CVPR.2015.7298794>.

- [56] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
- [57] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “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*, 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.
- [58] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2360–2367. doi: 10.1109/CVPR.2010.5539926.
- [59] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline),” in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, ser. Lecture Notes in Computer Science, vol. 11208, 2018, pp. 501–518. doi: 10.1007/978-3-030-01225-0_30.
- [60] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, “Learning discriminative features with multiple granularities for person re-identification,” in *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, 2018, pp. 274–282. doi: 10.1145/3240508.3240552.
- [61] L. Sun, J. Liu, Y. Zhu, and Z. Jiang, “Local to global with multi-scale attention network for person re-identification,” in *2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019*, 2019, pp. 2254–2258. doi: 10.1109/ICIP.2019.8803292.

- [62] X. Chen, C. Fu, Y. Zhao, *et al.*, “Salience-guided cascaded suppression network for person re-identification,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020, pp. 3297–3307. **doi:** 10.1109/CVPR42600.2020.00336.
- [63] A. Corrado, S. Barba, I. Carozzo, and S. Nardi, “Smart passenger center: Real-time optimization of urban public transport,” in *Proceedings of the Thirty-Sixth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2023, Clearwater Beach, FL, USA, May 14-17, 2023*, M. Franklin and S. A. Chun, Eds., 2023. **doi:** 10.32473/FLAIRS.36.133300.
- [64] J. Gu, K. Wang, H. Luo, *et al.*, “Msinet: Twins contrastive search of multi-scale interaction for object reid,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, IEEE, 2023, pp. 19 243–19 253. **doi:** 10.1109/CVPR52729.2023.01844. **url:** <https://doi.org/10.1109/CVPR52729.2023.01844>.
- [65] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, “Omni-scale feature learning for person re-identification,” in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 2019, pp. 3701–3711. **doi:** 10.1109/ICCV.2019.00380. **url:** <https://doi.org/10.1109/ICCV.2019.00380>.
- [66] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944. **doi:** 10.1109/CVPR.2017.106.
- [67] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, “Human semantic parsing for person re-identification,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*

- 2018, Salt Lake City, UT, USA, June 18-22, 2018, 2018, pp. 1062–1071. **doi:** 10.1109/CVPR.2018.00117.
- [68] R. Quispe and H. Pedrini, “Improved person re-identification based on saliency and semantic parsing with deep neural network models,” *Image and Vision Computing*, vol. 92, p. 103 809, 2019. **doi:** 10.1016/j.imavis.2019.07.009. **url:** <https://doi.org/10.1016/j.imavis.2019.07.009>.
- [69] L. Yang, W. Jia, S. Li, and Q. Song, “Deep learning technique for human parsing: A survey and outlook,” *Int. J. Comput. Vis.*, vol. 132, no. 8, pp. 3270–3301, 2024. **doi:** 10.1007/S11263-024-02031-9.
- [70] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” *arXiv preprint arXiv:1706.05587*, 2017. **doi:** 10.48550/arXiv.1706.05587. **url:** <https://doi.org/10.48550/arXiv.1706.05587>.
- [71] X. Li, L. Zhao, L. Wei, *et al.*, “Deepsaliency: Multi-task deep neural network model for salient object detection,” *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3919–3930, 2016. **doi:** 10.1109/TIP.2016.2579306. **url:** <https://doi.org/10.1109/TIP.2016.2579306>.
- [72] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, “Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 6757–6765. **doi:** 10.1109/CVPR.2017.715.
- [73] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian, “Visual attention methods in deep learning: An in-depth survey,” *Inf. Fusion*, vol. 108, p. 102 417, 2024. **doi:** 10.1016/J.INFFUS.2024.102417.
- [74] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” *IEEE Transactions on Pattern Analysis and Machine In-*

- telligence*, vol. 42, no. 8, pp. 2011–2023, 2020. **doi:** 10.1109/TPAMI.2019.2913372.
- [75] X. Wang, R. B. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018, pp. 7794–7803. **doi:** 10.1109/CVPR.2018.00813.
- [76] C. Li, S. Jin, E. Chang, S. Xuan, F. Liu, and D. Xu, “Scanet: Person re-identification with semantically consistent attention,” in *2020 Chinese Automation Congress (CAC)*, IEEE, 2020, pp. 3424–3428. **doi:** 10.1109/CAC51589.2020.9326801. **url:** <https://doi.org/10.1109/CAC51589.2020.9326801>.
- [77] X. Zhu, J. Qian, H. Wang, and P. Liu, “Curriculum enhanced supervised attention network for person re-identification,” *IEEE Signal Processing Letters*, vol. 27, pp. 1665–1669, 2020. **doi:** 10.1109/LSP.2020.3024794. **url:** <https://doi.org/10.1109/LSP.2020.3024794>.
- [78] V. Somers, A. Alahi, and C. D. Vleeschouwer, “Keypoint promptable re-identification,” in *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXIX*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., ser. Lecture Notes in Computer Science, vol. 15137, Springer, 2024, pp. 216–233. **doi:** 10.1007/978-3-031-72986-7_13. **url:** https://doi.org/10.1007/978-3-031-72986-7_13.
- [79] A. Benzine, M. E. A. Seddik, and J. Desmarais, “Deep miner: A deep and multi-branch network which mines rich and diverse features for person re-identification,” *arXiv preprint arXiv:2102.09321*, 2021.
- [80] R. Quispe and H. Pedrini, “Top-db-net: Top dropblock for activation enhancement in person re-identification,” in *25th International Conference on Pattern Recognition (ICPR 2020), Virtual Event / Milan, Italy, January 10-15, 2021*, 2020, pp. 2980–2987. **doi:** 10.1109/ICPR4

- 8806.2021.9412017. **url:** <https://doi.org/10.1109/ICPR48806.2021.9412017>.
- [81] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *9th International Conference on Learning Representations (ICLR 2021), Virtual Event, Austria, May 3-7, 2021*, 2021. **url:** <https://openreview.net/forum?id=YicbFdNTTy>.
- [82] C. Sharma, S. R. Kapil, and D. Chapman, “Person re-identification with a locally aware transformer,” *CoRR*, vol. abs/2106.03720, 2021. **doi:** <https://doi.org/10.48550/arXiv.2106.03720>. arXiv: 2106.03720. **url:** <https://arxiv.org/abs/2106.03720>.
- [83] B. Wu, C. Xu, X. Dai, *et al.*, “Visual transformers: Token-based image representation and processing for computer vision,” *arXiv preprint arXiv:2006.03677*, 2020.
- [84] W. Wang, E. Xie, X. Li, *et al.*, “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,” in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 2021, pp. 548–558. **doi:** 10.1109/ICCV48922.2021.00061.
- [85] H. Chen, Y. Wang, T. Guo, *et al.*, “Pre-trained image processing transformer,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2021, pp. 12 299–12 310. **doi:** 10.1109/CVPR46437.2021.01212.
- [86] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, “Transreid: Transformer-based object re-identification,” in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 2021, pp. 14 993–15 002. **doi:** 10.1109/ICCV48922.2021.01474.

- [87] Z. Zhang, D. He, S. Liu, B. Xiao, and T. S. Durrani, “Completed part transformer for person re-identification,” *IEEE Trans. Multim.*, vol. 26, pp. 2303–2313, 2024. **doi:** 10.1109/TMM.2023.3294816. **url:** <https://doi.org/10.1109/TMM.2023.3294816>.
- [88] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline),” in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., ser. Lecture Notes in Computer Science, vol. 11208, Springer, 2018, pp. 501–518. **doi:** 10.1007/978-3-030-01225-0_30. **url:** https://doi.org/10.1007/978-3-030-01225-0_30.
- [89] H. Chen, Y. Wang, Y. Shi, *et al.*, “Deep transfer learning for person re-identification,” in *Fourth IEEE International Conference on Multimedia Big Data, BigMM 2018, Xi’an, China, September 13-16, 2018*, 2018, pp. 1–5. **doi:** 10.1109/BIGMM.2018.8499067.
- [90] L. Wei, S. Zhang, W. Gao, and Q. Tian, “Person transfer gan to bridge domain gap for person re-identification,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 79–88. **doi:** 10.1109/CVPR.2018.00016. **url:** <https://doi.org/10.1109/CVPR.2018.00016>.
- [91] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, “Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018, pp. 994–1003. **doi:** 10.1109/CVPR.2018.00110.
- [92] Y. Chen, X. Zhu, and S. Gong, “Instance-guided context rendering for cross-domain person re-identification,” in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea*

- (South), October 27 - November 2, 2019, 2019, pp. 232–242. **doi:** 10.1109/ICCV.2019.00032.
- [93] X. Liu, H. Tan, X. Tong, J. Cao, and J. Zhou, “Feature preserving gan and multi-scale feature enhancement for domain adaption person re-identification,” *Neurocomputing*, vol. 364, pp. 108–118, 2019. **doi:** <https://doi.org/10.1016/j.neucom.2019.07.063>. **url:** <https://www.sciencedirect.com/science/article/pii/S0925231219310689>.
- [94] B. Sun and K. Saenko, “Deep CORAL: correlation alignment for deep domain adaptation,” in *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, 2016, pp. 443–450. **doi:** 10.1007/978-3-319-49409-8_35.
- [95] M. Wang and W. Deng, “Deep visual domain adaptation: A survey,” *Neurocomputing*, vol. 312, pp. 135–153, 2018. **doi:** <https://doi.org/10.1016/j.neucom.2018.05.083>. **url:** <https://www.sciencedirect.com/science/article/pii/S0925231218306684>.
- [96] J. Ho, A. Jain, and P. 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*, H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H.-T. Lin, Eds., 2020. **url:** <https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html>.
- [97] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021. **url:** <https://openreview.net/forum?id=St1giarCHLP>.

- [98] A. Asperti, S. Fiorilla, and L. Orsini, “A generative approach to person reidentification,” *Sensors*, vol. 24, no. 4, p. 1240, 2024. **doi:** 10.3390/S24041240. **url:** <https://doi.org/10.3390/s24041240>.
- [99] A. Asperti, L. Naldi, and S. Fiorilla, “An investigation of the domain gap in clip-based person re-identification,” *Sensors*, vol. 25, no. 2, p. 363, 2025. **doi:** 10.3390/s25020363. **url:** <https://doi.org/10.3390/s25020363>.
- [100] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5962–5979, 2022. **doi:** 10.1109/TPAMI.2021.3087709. **url:** <https://doi.org/10.1109/TPAMI.2021.3087709>.
- [101] Y. Zhai, X. Guo, Y. Lu, and H. Li, “In defense of the classification loss for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019, pp. 1526–1535. **doi:** 10.1109/CVPRW.2019.00194.
- [102] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, “Bag of tricks for image classification with convolutional neural networks,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 558–567. **doi:** 10.1109/CVPR.2019.00065.
- [103] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, “Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018, pp. 994–1003. **doi:** 10.1109/CVPR.2018.00110.
- [104] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, “Group consistent similarity learning via deep CRF for person re-identification,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*

- 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 8649–8658. **doi:** 10.1109/CVPR.2018.00902. **url:** http://openaccess.thecvf.com/content%5C_cvpr%5C_2018/html/Chen%5C_Group%5C_Consistent%5C_Similarity%5C_CVPR%5C_2018%5C_paper.html.
- [105] M. Wiecek, B. Rychalska, and J. Dabrowski, “On the unreasonable effectiveness of centroids in image retrieval,” in *Neural Information Processing - 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8-12, 2021, Proceedings, Part IV*, ser. Lecture Notes in Computer Science, vol. 13111, 2021, pp. 212–223. **doi:** 10.1007/978-3-030-92273-3_18.
- [106] S. Zhang, Q. Zhang, X. Wei, Y. Zhang, and Y. Xia, “Person re-identification with triplet focal loss,” *IEEE Access*, vol. 6, pp. 78 092–78 099, 2018. **doi:** 10.1109/ACCESS.2018.2884743. **url:** <https://ieeexplore.ieee.org/document/8558553>.
- [107] T. Si, Z. Zhang, and S. Liu, “Compact triplet loss for person re-identification in camera sensor networks,” *Ad Hoc Networks*, vol. 95, p. 101 984, 2019. **doi:** 10.1016/j.adhoc.2019.101984. **url:** <https://doi.org/10.1016/j.adhoc.2019.101984>.
- [108] D. Zhao, C. Chen, and D. Li, “Multi-stage attention and center triplet loss for person re-identification,” *Applied Intelligence*, vol. 52, no. 3, pp. 3077–3089, 2022. **doi:** 10.1007/s10489-021-02511-2. **url:** <https://doi.org/10.1007/s10489-021-02511-2>.
- [109] A. Alnissany and Y. Dayoub, “Modified centroid triplet loss for person re-identification,” *Journal of Big Data*, vol. 10, no. 1, p. 74, 2023. **doi:** <https://doi.org/10.1186/s40537-023-00753-0>.
- [110] W. Chen, X. Chen, J. Zhang, and K. Huang, “Beyond triplet loss: A deep quadruplet network for person re-identification,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*,

- Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 1320–1329. **doi:** 10.1109/CVPR.2017.145.
- [111] Z. Zhang, C. Lan, W. Zeng, Z. Chen, and S. Chang, “Beyond triplet loss: Meta prototypical n-tuple loss for person re-identification,” *IEEE Trans. Multim.*, vol. 24, pp. 4158–4169, 2022. **doi:** 10.1109/TMM.2021.3115451. **url:** <https://doi.org/10.1109/TMM.2021.3115451>.
- [112] H. Shi, Y. Yang, X. Zhu, *et al.*, “Embedding deep metric for person re-identification: A study against large variations,” in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., ser. Lecture Notes in Computer Science, vol. 9905, 2016, pp. 732–748. **doi:** 10.1007/978-3-319-46448-0_44.
- [113] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, “Working hard to know your neighbor’s margins: Local descriptor learning loss,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 4826–4837. **url:** <https://proceedings.neurips.cc/paper/2017/hash/831caa1b600f852b7844499430ecac17-Abstract.html>.
- [114] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the 37th International Conference on Machine Learning (ICML 2020), 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 1597–1607. **url:** <http://proceedings.mlr.press/v119/chen20j.html>.
- [115] X. Ni and E. Rahtu, “Flipreid: Closing the gap between training and inference in person re-identification,” in *9th European Workshop on Visual Information Processing, EUVIP 2021, Paris, France, June 23-*

- 25, 2021, A. Beghdadi, F. A. Cheikh, J. M. R. S. Tavares, *et al.*, Eds., 2021, pp. 1–6. **doi:** 10.1109/EUVIP50544.2021.9484010.
- [116] X. Zang, G. Li, W. Gao, and X. Shu, “Learning to disentangle scenes for person re-identification,” *Image Vis. Comput.*, vol. 116, p. 104330, 2021. **doi:** 10.1016/J.IMAVIS.2021.104330. **url:** <https://doi.org/10.1016/j.imavis.2021.104330>.
- [117] X. Shi, H. Liu, W. Shi, Z. Zhou, and Y. Li, “Boosting person re-identification with viewpoint contrastive learning and adversarial training,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. **doi:** 10.1109/ICASSP49357.2023.10095823.
- [118] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [119] Q. Wang, X. Qian, B. Li, Y. Fu, and X. Xue, “Rethinking person re-identification from a projection-on-prototypes perspective,” *arXiv preprint arXiv:2308.10717*, 2023.
- [120] A. Radford, J. W. Kim, C. Hallacy, *et al.*, “Learning transferable visual models from natural language supervision,” *CoRR*, 2021. arXiv: 2103.00020. **url:** <https://arxiv.org/abs/2103.00020>.
- [121] S. Yan, N. Dong, L. Zhang, and J. Tang, “Clip-driven fine-grained text-image person re-identification,” *IEEE Trans. Image Process.*, vol. 32, pp. 6032–6046, 2023. **doi:** 10.1109/TIP.2023.3327924. **url:** <https://doi.org/10.1109/TIP.2023.3327924>.
- [122] K. Kim, M. Kim, H. Kim, S. Park, and J. Paik, “Person re-identification method using text description through CLIP,” in *International Conference on Electronics, Information, and Communication, ICEIC 2023, Singapore, February 5-8, 2023*, IEEE, 2023, pp. 1–4. **doi:** 10.1109/ICEIC57457.2023.10049924. **url:** <https://doi.org/10.1109/ICEIC57457.2023.10049924>.

- [123] S. Li, L. Sun, and Q. Li, “CLIP-ReID: Exploiting vision-language model for image re-identification without concrete text labels,” in *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 2023, pp. 1405–1413. **doi:** 10.1609/AAAI.V37I1.25225.
- [124] C. Yu, X. Liu, Y. Wang, P. Zhang, and H. Lu, “TF-CLIP: learning text-free CLIP for video-based person re-identification,” in *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, M. J. Wooldridge, J. G. Dy, and S. Natarajan, Eds., AAAI Press, 2024, pp. 6764–6772. **doi:** 10.1609/AAAI.V38I7.28500.
- [125] Z. Shao, X. Zhang, C. Ding, J. Wang, and J. Wang, “Unified pre-training with pseudo texts for text-to-image person re-identification,” in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 2023, pp. 11 140–11 150. **doi:** 10.1109/ICCV51070.2023.01026.
- [126] Q. Yang, A. Wu, and W.-S. Zheng, “Person re-identification by contour sketch under moderate clothing change,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2029–2046, 2019. **doi:** 10.1109/TPAMI.2019.2960509.
- [127] Y. Huang, Q. Wu, J. Xu, Y. Zhong, and Z. Zhang, “Clothing status awareness for long-term person re-identification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 875–11 884. **doi:** 10.1109/ICCV48922.2021.01168.

- [128] Y. Ding, Y. Wu, A. Wang, T. Gong, and L. Zhang, “Disentangled body features for clothing change person re-identification,” *Multimedia Tools and Applications*, vol. 83, no. 27, pp. 69 693–69 714, 2024. **doi:** 10.1007/s11042-024-18440-4.
- [129] F. Wan, Y. Wu, X. Qian, Y. Chen, and Y. Fu, “When person re-identification meets changing clothes,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 3620–3628. **doi:** 10.1109/CVPRW50498.2020.00423.
- [130] H. Tu, C. Liu, Y. Peng, H. Xiong, and H. Wang, “Clothing-change person re-identification based on fusion of rgb modality and gait features,” *Signal, Image and Video Processing*, vol. 18, no. 3, pp. 2367–2376, 2024. **doi:** 10.1007/s11760-023-02913-4.
- [131] V. D. Nguyen, P. Mantini, and S. K. Shah, “Contrastive clothing and pose generation for cloth-changing person re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024, pp. 7541–7549. **doi:** 10.1109/cvprw63382.2024.00749.
- [132] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: Analysis and implementation,” *on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002. **doi:** 10.1109/TPAMI.2002.1017616. **url:** <https://doi.org/10.1109/TPAMI.2002.1017616>.
- [133] M. Ester, H. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, E. Simoudis, J. Han, and U. M. Fayyad, Eds., AAAI Press, 1996, pp. 226–231. **url:** <http://www.aaai.org/Library/KDD/1996/kdd96-037.php>.

- [134] H. Fan, L. Zheng, C. Yan, and Y. Yang, “Unsupervised person re-identification: Clustering and fine-tuning,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 4, pp. 1–18, 2018. **doi:** 10.1145/3243316. **url:** <https://doi.org/10.1145/3243316>.
- [135] X. Lin, P. Ren, C.-H. Yeh, L. Yao, A. Song, and X. Chang, “Unsupervised person re-identification: A systematic survey of challenges and solutions,” *arXiv preprint arXiv:2109.06057*, 2021.
- [136] S. Wang, L. Zhang, W. Chen, F. Wang, and H. Li, “Refining pseudo labels for unsupervised domain adaptive re-identification,” *Knowl. Based Syst.*, vol. 242, p. 108 336, 2022. **doi:** 10.1016/J.KNOSYS.2022.108336. **url:** <https://doi.org/10.1016/j.knosys.2022.108336>.
- [137] P. Zhao, L. Huang, W. Zhang, X. Li, and Z. Wei, “Exploiting reliable pseudo-labels for unsupervised domain adaptive person re-identification,” *Neurocomputing*, vol. 492, pp. 581–592, 2022. **doi:** <https://doi.org/10.1016/j.neucom.2021.12.050>.
- [138] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 1195–1204. **url:** <https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html>.
- [139] B. T. Polyak and A. B. Juditsky, “Acceleration of stochastic approximation by averaging,” *SIAM Journal on Control and Optimization*, vol. 30, no. 4, pp. 838–855, 1992. **doi:** 10.1137/0330046. **url:** <https://doi.org/10.1137/0330046>.
- [140] Y. Ge, D. Chen, and H. Li, “Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification,” in *8th International Conference on Learning Representations, ICLR*

- 2020, Addis Ababa, Ethiopia, April 26-30, 2020, 2020. **url:** <https://openreview.net/forum?id=rJln0hVYPS>.
- [141] Y. Guo, F. Feng, X. Hao, and X. Chen, “Jac-net: Joint learning with adaptive exploration and concise attention for unsupervised domain adaptive person re-identification,” *Neurocomputing*, vol. 483, pp. 262–274, 2022. **doi:** 10.1016/j.neucom.2022.02.010. **url:** <https://doi.org/10.1016/j.neucom.2022.02.010>.
- [142] D. Cheng, J. Zhou, N. Wang, and X. Gao, “Hybrid dynamic contrast and probability distillation for unsupervised person re-id,” *on Image Processing*, vol. 31, pp. 3334–3346, 2022. **doi:** 10.1109/TIP.2022.3169693.
- [143] J. Ding and X. Zhou, “Learning feature fusion for unsupervised domain adaptive person re-identification,” in *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE, 2022, pp. 2613–2619. **doi:** 10.1109/ICPR56361.2022.9956264. **url:** <https://doi.org/10.1109/ICPR56361.2022.9956264>.
- [144] S. Wang, Y. Cui, and Y. Kang, “Learning multiple granularity features for unsupervised person re-identification,” in *IEEE International Conference on Multimedia and Expo, ICME 2022, Taipei, Taiwan, July 18-22, 2022*, IEEE, 2022, pp. 1–6. **doi:** 10.1109/ICME52920.2022.9859983. **url:** <https://doi.org/10.1109/ICME52920.2022.9859983>.
- [145] Y. Ge, F. Zhu, D. Chen, R. Zhao, and H. Li, “Self-paced contrastive learning with hybrid memory for domain adaptive object re-id,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020, pp. 11 309–11 321. **url:** <https://proceedings.neurips.cc/paper/2020/hash/821fa74b50ba3f7c ba1e6c53e8fa6845-Abstract.html>.

- [146] X. Song, J. Liu, and Z. Jin, “Dual prototype contrastive learning with fourier generalization for domain adaptive person re-identification,” *Knowledge-Based Systems*, vol. 256, p. 109851, 2022. **doi:** <https://doi.org/10.1016/j.knosys.2022.109851>.
- [147] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. S. Huang, “Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification,” in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 2019, pp. 6111–6120. **doi:** [10.1109/ICCV.2019.00621](https://doi.org/10.1109/ICCV.2019.00621).
- [148] P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe, “Curriculum learning: A survey,” *International Journal of Computer Vision*, vol. 130, no. 6, pp. 1526–1565, 2022. **doi:** [10.1007/s11263-022-01611-x](https://doi.org/10.1007/s11263-022-01611-x). **url:** <https://doi.org/10.1007/s11263-022-01611-x>.
- [149] Z. Guo, B. Ma, H. Chang, and X. Chen, “Gradual domain adaptation with sample transferability exploitation for person re-identification,” in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 2022, pp. 1–6. **doi:** [10.1109/ICME52920.2022.9859861](https://doi.org/10.1109/ICME52920.2022.9859861).
- [150] S. Chen, L. Qiu, Z. Tian, Y. Yan, D.-H. Wang, and S. Zhu, “Mtnet: Mutual tri-training network for unsupervised domain adaptation on person re-identification,” *Journal of Visual Communication and Image Representation*, vol. 90, p. 103749, 2023. **doi:** [10.1016/j.jvcir.2022.103749](https://doi.org/10.1016/j.jvcir.2022.103749). **url:** <https://doi.org/10.1016/j.jvcir.2022.103749>.
- [151] Z. Liu, B. Liu, Z. Zhao, Q. Chu, and N. Yu, “Dual-uncertainty guided curriculum learning and part-aware feature refinement for domain adaptive person re-identification,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. **doi:** [10.1109/ICASSP49357.2023.10097020](https://doi.org/10.1109/ICASSP49357.2023.10097020).

- [152] D. Fu, D. Chen, J. Bao, *et al.*, “Unsupervised pre-training for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2021, pp. 14 750–14 759. **doi:** 10.1109/CVPR46437.2021.01451.
- [153] M. Li, C.-G. Li, and J. Guo, “Cluster-guided asymmetric contrastive learning for unsupervised person re-identification,” *on Image Processing*, vol. 31, pp. 3606–3617, 2022. **doi:** 10.1109/TIP.2022.3173163.
- [154] J. Li, M. Wang, and X. Gong, “Transformer based multi-grained features for unsupervised person re-identification,” in *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV 2023 - Workshops, Waikoloa, HI, USA, January 3-7, 2023*, IEEE, 2023, pp. 1–9. **doi:** 10.1109/WACVW58289.2023.00009. **url:** <https://doi.org/10.1109/WACVW58289.2023.00009>.
- [155] H. Luo, P. Wang, Y. Xu, *et al.*, “Self-supervised pre-training for transformer-based person re-identification,” *arXiv preprint*, 2021. arXiv: 2111.12084. **url:** <https://arxiv.org/abs/2111.12084>.
- [156] B. Pang, D. Zhai, J. Jiang, and X. Liu, “Fully unsupervised person re-identification via selective contrastive learning,” *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 18, no. 2, pp. 1–15, 2022. **doi:** 10.1145/3485061.
- [157] W. Li, C. Chen, and K. Huang, “Absorb and repel: Pseudo-label refinement for intra-camera supervised person re-identification,” *on Artificial Intelligence*, pp. 1–14, 2023. **doi:** 10.1109/TAI.2023.3327671.
- [158] T. Liu, Y. Lin, and B. Du, “Unsupervised person re-identification with stochastic training strategy,” *on Image Processing*, vol. 31, pp. 4240–4250, 2022. **doi:** 10.1109/TIP.2022.3181811.
- [159] G. Zhang, H. Zhang, W. Lin, A. K. Chandran, and X. Jing, “Camera contrast learning for unsupervised person re-identification,” *on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 4096–4107, 2023. **doi:** 10.1109/TCSVT.2023.3240001.

- [160] Y. Liu, M. Qi, Q. Wu, Y. Yang, X. Li, and J. Zhang, “Camera proxy based contrastive learning with hard sampling for unsupervised person re-identification,” in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2023, pp. 2423–2428. **doi:** 10.1109/ICME55011.2023.00413.
- [161] M. Wang, J. Li, B. Lai, X. Gong, and X.-S. Hua, “Offline-online associated camera-aware proxies for unsupervised person re-identification,” *on Image Processing*, vol. 31, pp. 6548–6561, 2022. **doi:** 10.1109/TIP.2022.3213193.
- [162] X. Zhang, D. Li, Z. Wang, *et al.*, “Implicit sample extension for unsupervised person re-identification,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 7359–7368. **doi:** 10.1109/CVPR52688.2022.00722.
- [163] Z. Dai, G. Wang, W. Yuan, S. Zhu, and P. Tan, “Cluster contrast for unsupervised person re-identification,” in *Computer Vision - ACCV 2022 - 16th Asian Conference on Computer Vision, Macao, China, December 4-8, 2022, Proceedings, Part VI*, L. Wang, J. Gall, T. Chin, I. Sato, and R. Chellappa, Eds., ser. Lecture Notes in Computer Science, vol. 13846, 2022, pp. 319–337. **doi:** 10.1007/978-3-031-26351-4_20.
- [164] L. Lan, X. Teng, J. Zhang, X. Zhang, and D. Tao, “Learning to purification for unsupervised person re-identification,” *on Image Processing*, vol. 32, pp. 3338–3353, 2023. **doi:** 10.1109/TIP.2023.3278860.
- [165] Y. Chen, Z. Fan, Z. Chen, and Y. Zhu, “Ca-jaccard: Camera-aware jaccard distance for person re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 17-21, 2024, Seattle, WA, 2024*, pp. 17 532–17 541. **doi:** 10.1109/CVPR52733.2024.01660. **url:** <https://doi.org/10.1109/CVPR52733.2024.01660>.

- [166] Y. Feng, S. Zhao, Y. Zhang, Y. Liu, S. Zhu, and S. Coleman, “Noise-tolerant learning with silhouette coefficient for unsupervised person re-identification,” in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 2022, pp. 1–6. doi: 10.1109/ICME52920.2022.9859824.
- [167] J. Yin, X. Zhang, Z. Ma, J. Guo, and Y. Liu, “A real-time memory updating strategy for unsupervised person re-identification,” *on Image Processing*, vol. 32, pp. 2309–2321, 2023. doi: 10.1109/TIP.2023.3266166.
- [168] Y. Liu, H. Ge, Z. Wang, Y. Hou, and M. Zhao, “Discriminative identity-feature exploring and differential aware learning for unsupervised person re-identification,” *on Multimedia*, vol. 26, pp. 623–636, 2024. doi: 10.1109/TMM.2023.3268369.
- [169] H. Zhang, G. Zhang, Y. Chen, and Y. Zheng, “Global relation-aware contrast learning for unsupervised person re-identification,” *on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8599–8610, 2022. doi: 10.1109/TCSVT.2022.3194084.
- [170] X. Bai, A. Wang, C. Zhang, and H. Hu, “Cross-domain person re-identification based on normalized ibn-net,” *IEEE Access*, vol. 12, pp. 54 220–54 228, 2024. doi: 10.1109/ACCESS.2024.3387478. url: <https://doi.org/10.1109/ACCESS.2024.3387478>.
- [171] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, “Simam: A simple, parameter-free attention module for convolutional neural networks,” in *Proceedings of the 38th International Conference on Machine Learning (ICML 2021), 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 11 863–11 874. url: <http://proceedings.mlr.press/v139/yang21o.html>.
- [172] J. Peng, G. Jiang, and H. Wang, “Adaptive memorization with group labels for unsupervised person re-identification,” *on Circuits and Sys-*

- tems for Video Technology*, vol. 33, no. 10, pp. 5802–5813, 2023. **doi:** 10.1109/TCSVT.2023.3258917.
- [173] H. Chen, F. Bremond, N. Sebe, and S. Zhang, “Anti-forgetting adaptation for unsupervised person re-identification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 2, pp. 1056–1072, 2025. **doi:** 10.1109/TPAMI.2024.3490777.
- [174] K. Xu, X. Zou, and J. Zhou, “Lstkc: Long short-term knowledge consolidation for lifelong person re-identification,” in *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, vol. 38, AAAI Press, 2024, pp. 16 202–16 210. **doi:** 10.1609/aaai.v38i14.29554. **url:** <https://doi.org/10.1609/aaai.v38i14.29554>.
- [175] Z. Huang, Z. Zhang, C. Lan, *et al.*, “Lifelong unsupervised domain adaptive person re-identification with coordinated anti-forgetting and adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, June 18–24, 2022*, IEEE, 2022, pp. 14 268–14 277. **doi:** 10.1109/CVPR52688.2022.01389. **url:** <https://doi.org/10.1109/CVPR52688.2022.01389>.
- [176] W. Chen, X. Xu, J. Jia, *et al.*, “Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17–24, 2023*, 2023, pp. 15 050–15 061. **doi:** 10.1109/CVPR52729.2023.01445.
- [177] Y. Ci, Y. Wang, M. Chen, *et al.*, “Unihcp: A unified model for human-centric perceptions,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17–24, 2023*, 2023, pp. 17 840–17 852. **doi:** 10.1109/CVPR52729.2023.01711.

- [178] G. Wang, J. Lai, P. Huang, and X. Xie, “Spatial-temporal person re-identification,” in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 2019, pp. 8933–8940. **doi:** 10.1609/AAAI.V33I01.33018933.
- [179] F. Herzog, X. Ji, T. Teepe, S. Hörmann, J. Gilg, and G. Rigoll, “Lightweight multi-branch network for person re-identification,” in *2021 IEEE International Conference on Image Processing, ICIP 2021, Anchorage, AK, USA, September 19-22, 2021*, 2021, pp. 1129–1133. **doi:** 10.1109/ICIP42928.2021.9506733.
- [180] J.-J. Wu, K.-H. Chang, and I.-C. Lin, “Generalizable person re-identification with part-based multi-scale network,” *Multimedia Tools and Applications*, vol. 82, no. 25, pp. 38 639–38 666, 2023. **doi:** 10.1007/s11042-023-14718-1. **url:** <https://doi.org/10.1007/s11042-023-14718-1>.
- [181] Y. Huang, P. Peng, Y. Jin, Y. Li, and J. Xing, “Domain adaptive attention learning for unsupervised person re-identification,” in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 2020, pp. 11 069–11 076. **doi:** 10.1609/AAAI.V34I07.6762. **url:** <https://doi.org/10.1609/aaai.v34i07.6762>.
- [182] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, “Learning invariance from generated variance for unsupervised person re-identification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7494–7508, 2023. **doi:** 10.1109/TPAMI.2022.3226866.

- [183] H. Fan, L. Zheng, C. Yan, and Y. Yang, “Unsupervised person re-identification: Clustering and fine-tuning,” *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 14, no. 4, 83:1–83:18, 2018. **doi:** 10.1145/3243316. **url:** <https://doi.org/10.1145/3243316>.
- [184] S. Wang, L. Zhang, W. Chen, F. Wang, and H. Li, “Refining pseudo labels for unsupervised domain adaptive re-identification,” *Knowl. Based Syst.*, vol. 242, p. 108336, 2022. **doi:** 10.1016/J.KNOSYS.2022.108336. **url:** <https://doi.org/10.1016/j.knosys.2022.108336>.
- [185] T. Yan, K. Zhu, H. Guo, G. Zhu, M. Tang, and J. Wang, “Plug-and-play pseudo label correction network for unsupervised person re-identification,” *CoRR*, vol. abs/2206.06607, 2022. **doi:** 10.48550/ARXIV.2206.06607. arXiv: 2206.06607. **url:** <https://doi.org/10.48550/arXiv.2206.06607>.
- [186] A. Asperti and V. Tonelli, “Comparing the latent space of generative models,” *Neural Computing & Applications*, vol. 35, pp. 3155–3172, 2023. **doi:** 10.1007/s00521-022-07890-2. **url:** <https://link.springer.com/article/10.1007/s00521-022-07890-2>.
- [187] A. Asperti, G. Colasuonno, and A. Guerra, “Portrait reification with generative diffusion models,” *Applied Sciences*, vol. 13, no. 11, 2023. **doi:** 10.3390/app13116487. **url:** <https://www.mdpi.com/2076-3417/13/11/6487>.
- [188] P. Dhariwal and A. Q. 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*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 8780–8794. **url:** <https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html>.

- [189] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” *ArXiv*, vol. abs/2204.06125, 2022.
- [190] C. Saharia, W. Chan, S. Saxena, *et al.*, “Photorealistic text-to-image diffusion models with deep language understanding,” *CoRR*, vol. abs/2205.11487, 2022. **doi:** 10.48550/arXiv.2205.11487. **arXiv:** 2205.11487. **url:** <https://doi.org/10.48550/arXiv.2205.11487>.
- [191] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, “Video diffusion models,” *arXiv:2204.03458*, 2022.
- [192] A. Asperti, F. Merizzi, A. Paparella, G. Pedrazzi, M. Angelinelli, and S. Colamonaco, “Precipitation nowcasting with generative diffusion models,” *Applied Intelligence*, vol. 55, no. 2, pp. 1–21, 2025. **doi:** 10.1007/s10489-024-06048-y. **url:** <https://doi.org/10.1007/s10489-024-06048-y>.
- [193] A. Asperti, D. Evangelista, S. Marro, and F. Merizzi, “Image embedding for denoising generative models,” *Artificial Intelligence Review*, vol. 56, pp. 14 511–14 533, 2023. **doi:** 10.1007/s10462-023-10504-5. **url:** <https://doi.org/10.1007/s10462-023-10504-5>.
- [194] A. Asperti, G. Colasuonno, and A. Guerra, “Head rotation in denoising diffusion models,” *arXiv preprint arXiv:2308.06057*, 2023. **url:** <https://arxiv.org/abs/2308.06057>.
- [195] J. Ho and T. Salimans, “Classifier-free diffusion guidance,” *CoRR*, vol. abs/2207.12598, 2022. **doi:** 10.48550/arXiv.2207.12598. **arXiv:** 2207.12598. **url:** <https://doi.org/10.48550/arXiv.2207.12598>.
- [196] A. Odena, C. Olah, and J. Shlens, “Conditional image synthesis with auxiliary classifier gans,” in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, ser. Proceedings of Machine Learning Research, vol. 70, 2017, pp. 2642–2651. **url:** <http://proceedings.mlr.press/v70/odena17a.html>.

- [197] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.
- [198] A. Vaswani, N. Shazeer, N. Parmar, *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*, 2017, pp. 5998–6008. **url:** <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547de91fbd053c1c4a845aa-Abstract.html>.
- [199] D. Kingma, T. Salimans, B. Poole, and J. Ho, “Variational diffusion models,” *Advances in neural information processing systems*, vol. 34, pp. 21 696–21 707, 2021.
- [200] A. Q. Nichol and P. Dhariwal, “Improved denoising diffusion probabilistic models,” in *Proceedings of the 38th International Conference on Machine Learning (ICML 2021), Proceedings of Machine Learning Research*, ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 8162–8171. **url:** <https://proceedings.mlr.press/v139/nichol21a.html>.
- [201] A. Asperti and M. Trentin, “Balancing reconstruction error and kullback-leibler divergence in variational autoencoders,” *IEEE Access*, vol. 8, pp. 199 440–199 448, 2020. **doi:** 10.1109/ACCESS.2020.3034828.
- [202] A. Asperti, D. Evangelista, and E. L. Piccolomini, “A survey on variational autoencoders from a green AI perspective,” *SN Comput. Sci.*, vol. 2, no. 4, p. 301, 2021. **doi:** 10.1007/s42979-021-00702-9. **url:** <https://doi.org/10.1007/s42979-021-00702-9>.
- [203] B. Dai and D. P. Wipf, “Diagnosing and enhancing vae models,” in *Seventh International Conference on Learning Representations (ICLR 2019), May 6-9, New Orleans*, 2019. **url:** <https://openreview.net/forum?id=B1e0X3C9tQ>.

- [204] A. Asperti, “Variance loss in variational autoencoders,” in *Machine Learning, Optimization, and Data Science – 6th International Conference, LOD 2020, Siena, Italy, September 10-13, 2020, Proceedings*, ser. Lecture Notes in Computer Science, vol. 12565, Springer, 2020, pp. 297–308. **doi:** 10.1007/978-3-030-64583-0_28. **url:** https://doi.org/10.1007/978-3-030-64583-0_28.
- [205] A. Yadav and D. K. Vishwakarma, “Deep learning algorithms for person re-identification: State-of-the-art and research challenges,” *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 22 005–22 054, 2023. **doi:** 10.1007/s11042-023-16286-w.
- [206] Y.-J. Li, C.-S. Lin, Y.-B. Lin, and Y.-C. F. Wang, “Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7918–7928. **doi:** 10.1109/ICCV.2019.00801.
- [207] H. Li, Z. Kuang, Z. Yu, and J. Luo, “Structure alignment of attributes and visual features for cross-dataset person re-identification,” *Pattern Recognit.*, vol. 106, p. 107 414, 2020. **doi:** 10.1016/J.PATCOG.2020.107414. **url:** <https://doi.org/10.1016/j.patcog.2020.107414>.
- [208] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17, Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010.
- [209] R. Zhang, “Making convolutional networks shift-invariant again,” *CoRR*, vol. abs/1904.11486, 2019. arXiv: 1904.11486. **url:** <http://arxiv.org/abs/1904.11486>.
- [210] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022. **doi:** 10.1007/s11263-022-01653-1.

- [211] R. Quan, B. Xu, and D. Liang, “Discriminatively unsupervised learning person re-identification via considering complicated images,” *Sensors*, vol. 23, no. 6, 2023. **doi:** 10.3390/s23063259. **url:** <https://www.mdpi.com/1424-8220/23/6/3259>.
- [212] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 13 001–13 008, 2020. **doi:** 10.1609/aaai.v34i07.7000. **url:** <https://ojs.aaai.org/index.php/AAAI/article/view/7000>.
- [213] P. Marchwica, M. Jamieson, and P. Siva, “An evaluation of deep cnn baselines for scene-independent person re-identification,” in *2018 15th Conference on Computer and Robot Vision (CRV)*, 2018, pp. 297–304. **doi:** 10.1109/CRV.2018.00049.
- [214] T. Xiao, H. Li, W. Ouyang, and X. Wang, “Learning deep feature representations with domain guided dropout for person re-identification,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1249–1258. **doi:** 10.1109/CVPR.2016.140.
- [215] V. D. Nguyen, P. Mantini, and S. K. Shah, “Contrastive clothing and pose generation for cloth-changing person re-identification,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, IEEE, 2024, pp. 7541–7549. **doi:** 10.1109/CVPRW63382.2024.00749. **url:** <https://doi.org/10.1109/CVPRW63382.2024.00749>.
- [216] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, “Mot16: A benchmark for multi-object tracking,” *arXiv:1603.00831 [cs]*, 2016. **url:** <http://arxiv.org/abs/1603.00831>.
- [217] G. Maggolino, A. Ahmad, J. Cao, and K. Kitani, “Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification,” in *IEEE International Conference on Image Processing, ICIP 2023, Kuala Lumpur, Malaysia, October 8-11, 2023*, IEEE, 2023, pp. 3025–3029. **doi:** 10.1

- 109/ICIP49359.2023.10222576. **url:** <https://doi.org/10.1109/ICIP49359.2023.10222576>.
- [218] M. Kosela, J. Aszyk, M. Jarek, J. Klimek, and T. Prokop, “Tracking of spermatozoa by yolov5 detection and strongsort with osnet tracker,” in *Working Notes Proceedings of the MediaEval 2022 Workshop, Bergen, Norway and Online, 12-13 January 2023*, S. Hicks, A. G. S. de Herrera, J. Langguth, *et al.*, Eds., ser. CEUR Workshop Proceedings, vol. 3583, CEUR-WS.org, 2022. **url:** <https://ceur-ws.org/Vol-3583/paper41.pdf>.
- [219] J. Seidenschwarz, G. Brasó, V. C. Serrano, I. Elezi, and L. Leal-Taixé, “Simple cues lead to a strong multi-object tracker,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, IEEE, 2023, pp. 13 813–13 823. **doi:** 10.1109/CVPR52729.2023.01327. **url:** <https://doi.org/10.1109/CVPR52729.2023.01327>.
- [220] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, “Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization,” *arXiv preprint arXiv:1905.05172*, 2019.
- [221] W. E. Lorensen and H. E. Cline, “Marching cubes: A high resolution 3d surface construction algorithm,” in *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1987, Anaheim, California, USA, July 27-31, 1987*, M. C. Stone, Ed., ACM, 1987, pp. 163–169. **doi:** 10.1145/37401.37422. **url:** <https://doi.org/10.1145/37401.37422>.
- [222] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, “Fastreid: A pytorch toolbox for general instance re-identification,” in *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, A. El-Saddik, T. Mei, R. Cucchiara, *et al.*, Eds., ACM, 2023, pp. 9664–

9667. **doi:** 10.1145/3581783.3613460. **url:** <https://doi.org/10.1145/3581783.3613460>.
- [223] Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person re-identification with k-reciprocal encoding,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, IEEE Computer Society, 2017, pp. 3652–3661. **doi:** 10.1109/CVPR.2017.389. **url:** <https://doi.org/10.1109/CVPR.2017.389>.
- [224] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Association for Computational Linguistics, 2019, pp. 4171–4186. **doi:** 10.18653/V1/N19-1423. **url:** <https://doi.org/10.18653/v1/n19-1423>.
- [225] A. Alsehaim and T. P. Breckon, “Vid-trans-reid: Enhanced video transformers for person re-identification,” in *33rd British Machine Vision Conference 2022 (BMVC 2022), London, UK, November 21-24, 2022*, BMVA Press, 2022, p. 342. **url:** <https://bmvc2022.mpi-inf.mpg.de/342/>.
- [226] J. Wu, L. He, W. Liu, *et al.*, “Cavit: Contextual alignment vision transformer for video object re-identification,” in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIV*, ser. Lecture Notes in Computer Science, vol. 13674, Springer, 2022, pp. 549–566. **doi:** 10.1007/978-3-031-19781-9_32. **url:** https://doi.org/10.1007/978-3-031-19781-9_32.

- [227] J. Lee, Y. Lee, J. Kim, A. R. Kosiorek, S. Choi, and Y. W. Teh, “Set transformer: A framework for attention-based permutation-invariant neural networks,” in *Proceedings of the 36th International Conference on Machine Learning (ICML 2019), Long Beach, California, USA, 9-15 June 2019*, ser. Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 3744–3753. **url:** <http://proceedings.mlr.press/v97/lee19d.html>.
- [228] M. Kim, M. Cho, H. Lee, and S. Lee, “Spatio-temporal feature-level augmentation vision transformer for video-based person re-identification,” *Pattern Recognit.*, vol. 168, p. 111 813, 2025. **doi:** 10.1016/J.PATCOG.2025.111813. **url:** <https://doi.org/10.1016/j.patcog.2025.111813>.
- [229] X. Ma, J. Yi, and J. Gall, “Pet-keystation: Parameter-efficient transformer with keypoint-guided spatial-temporal aggregation for video-based person re-identification,” in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, IEEE, 2025, pp. 1–5. **doi:** 10.1109/ICASSP49660.2025.10888219. **url:** <https://doi.org/10.1109/ICASSP49660.2025.10888219>.
- [230] R. G. Brown and P. Y. C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering: With MATLAB Exercises and Solutions*, 3rd. New York, NY: Wiley, 1997.
- [231] J. Luiten, A. Osep, P. Dendorfer, *et al.*, “HOTA: A higher order metric for evaluating multi-object tracking,” *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 548–578, 2021. **doi:** 10.1007/S11263-020-01375-2. **url:** <https://doi.org/10.1007/s11263-020-01375-2>.
- [232] T. Stanczyk and F. Bremond, “Does re-id really help in multi-object tracking?” In *Lecture Notes of TICMI, Vol. 27*, Tbilisi, Georgia: Tbilisi University Press, 2026, pp. 1–15. **url:** <https://inria.hal.science/hal-05462063v1>.

- [233] J. Willemson, “Fifty shades of personal data - partial re-identification and GDPR,” in *Privacy Technologies and Policy - 10th Annual Privacy Forum, APF 2022, Warsaw, Poland, June 23-24, 2022, Proceedings*, A. Gryszczyńska, P. Polanski, N. Gruschka, K. Rannenberg, and M. Adamczyk, Eds., ser. Lecture Notes in Computer Science, vol. 13279, Springer, 2022, pp. 88–96. **doi:** 10.1007/978-3-031-07315-1_6. **url:** https://doi.org/10.1007/978-3-031-07315-1%5C_6.
- [234] B. Kim, S. Jeong, B. Cho, and J. Chung, “AI governance in the context of the EU AI act,” *IEEE Access*, vol. 13, pp. 144 126–144 142, 2025. **doi:** 10.1109/ACCESS.2025.3598023. **url:** <https://doi.org/10.1109/ACCESS.2025.3598023>.
- [235] C. C. Outeda, “The eu’s AI act: A framework for collaborative governance,” *Internet Things*, vol. 27, p. 101 291, 2024. **doi:** 10.1016/J.IOT.2024.101291. **url:** <https://doi.org/10.1016/j.iot.2024.101291>.
- [236] R. M. de Carvalho, C. D. Prete, Y. S. Martín, *et al.*, “Protecting citizens’ personal data and privacy: Joint effort from GDPR EU cluster research projects,” *SN Comput. Sci.*, vol. 1, no. 4, p. 217, 2020. **doi:** 10.1007/S42979-020-00218-8. **url:** <https://doi.org/10.1007/s42979-020-00218-8>.