



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN
BENI CULTURALI E AMBIENTALI

Ciclo 38

Settore Concorsuale: 01/B1 - INFORMATICA

Settore Scientifico Disciplinare: INF/01 - INFORMATICA

MULTI-TASK DEEP LEARNING FOR CHINESE PORCELAIN CLASSIFICATION:
ADDRESSING DATA SCARCITY THROUGH SYNTHETIC IMAGE GENERATION
AND TRANSFER LEARNING

Presentata da: Ziyao Ling

Coordinatore Dottorato

Donatella Restani

Supervisore

Silvia Mirri

Co-supervisore

Giovanni Delnevo

Esame finale anno 2026

**This dissertation is submitted for the degree of
Doctor of Philosophy in Beni Culturali e Ambientali
May 2026**

We shape our tools and
thereafter they shape us

—Marshall McLuhan

The past is never dead. It's
not even past

—William Faulkner

Acknowledgements

This thesis means more than just the end of my Ph.D journey, but a milestone in more than three years of work at Bologna University and specifically within the department of computer science and engineering.

First and foremost, I would like to express my deepest gratitude to my supervisor and co-supervisor, Prof. Silvia Mirri and Prof. Giovanni Delnevo, for their expertise, understanding, and patience added considerably to my graduate experience. Their guidance helped me throughout the time of research and writing of this thesis. They have made this PhD project possible and allowed me to navigate at my own pace while consistently guiding me forward.

My sincere thanks to Prof. Paola Salomoni for providing me the valuable guidance. Without her precious support, it would not have been possible to conduct this research.

I would like to acknowledge the financial support from China Scholarship Council (CSC) that funded parts of the research discussed in this thesis.

Finally, nobody has been more important to me in the pursuit of this project than the members of my family. I would like to thank my parents, whose love and guidance are with me in whatever I pursue. And my friends who have been extremely supportive of me throughout this entire process and have made countless sacrifices to help me get to this point.

Abstract

Chinese porcelain holds immense historical and cultural value, making its accurate classification essential for archaeological research and cultural heritage preservation. Traditional classification methods rely heavily on expert analysis, which is time-consuming, subjective, and difficult to scale. This thesis investigates the application of deep learning techniques to automate and assist the multi-task identification of Song and Yuan dynasty porcelains, addressing the critical constraint of limited annotated data in archaeological contexts. Through systematic experimentation with transfer learning, synthetic data generation, and multi-task learning frameworks, we develop a comprehensive classification system targeting four essential attributes: dynasty attribution, kiln provenance, glaze classification, and vessel typology. Explainable AI analysis reveals that the multi-task CNN model learns archaeologically meaningful features. These findings align with traditional identification practices, validating the model's decision-making process. By demonstrating that machine learning can effectively encode expert knowledge while handling data constraints typical of cultural heritage domains, this thesis establishes a reproducible framework applicable to other archaeological materials, ultimately bridging computational methods with cultural heritage.

Contents

List of Figures	viii
List of Tables	x
1 Introduction	1
1.1 Research Background	1
1.2 Research Questions	3
1.3 Outline	7
2 Machine Learning for Porcelain Classification: A Systematic Review	9
2.1 Introduction and Background	10
2.2 Methodology	12
2.2.1 PRISMA Methodology	12
2.2.2 Data Collection and Processing	13
2.2.3 Critical Assessment	15
2.2.4 Extraction of Relevant Fields	15
2.3 Datasets for Porcelain Identification(RQ1.1)	15
2.3.1 Datasets for archaeological ceramics identification	15
2.3.2 Type, country, and period of ceramics	17
2.3.3 Data source and sample count of ceramic	22
2.3.4 Data Scarcity and Future Directions	23
2.4 Image Processing Methods for Porcelain (RQ1.2)	24
2.4.1 Overview of Image Processing Methods	24
2.4.2 Focus on Feature Extraction, Image Segmentation, and Image Enhancement	26
2.4.3 Evolving Landscape of Image Processing Methods	28
2.5 Machine Learning Algorithms for Porcelain Classification (RQ1.3) . . .	30
2.5.1 Application of ML Algorithms in Archaeological Ceramic Images	30
2.5.2 Performance Evaluation	32
2.6 Discussion and Conclusion	33

3	Evaluating CNN Effectiveness in Multi-Task & Multi-Label Porcelain Classification	39
3.1	Introduction and Background	40
3.2	Dataset and Label Configuration	41
3.2.1	Data collection and collation stage	41
3.2.2	Historical and archaeological context	42
3.2.3	Hierarchical Label System and Historical Context	59
3.2.4	Train–Validation–Test Partitioning	61
3.2.5	Quantitative Combination-Imbalance Analysis	62
3.2.6	Image Pre-processing and Data-Augmentation	65
3.3	CNN Architecture and Multi-Task Learning Frameworks	67
3.3.1	Custom CNN Architecture	67
3.3.2	Multi-Task Learning Framework	69
3.3.3	Transfer Learning Models	72
3.3.4	Loss Function and Optimization	75
3.3.5	Training Strategy	79
3.3.6	Experimental Set-up	81
3.4	Results and Analysis	82
3.4.1	Quantitative Performance Evaluation (RQ2.1,RQ2.2)	82
3.4.2	Task-Specific Analysis (RQ2.1)	88
3.4.3	Error Analysis and Failure Cases (RQ2.1)	92
3.4.4	Model Interpretability Analysis (RQ2.3)	96
3.5	Discussion and Conclusion	102
3.5.1	Transfer Learning Superiority and Model Selection Insights	102
3.5.2	Task-Specific Insights and Archaeological Validation	103
3.5.3	Implications of the Sample Size Paradox	104
3.5.4	Multi-Task Learning Advantages and Limitations	104
3.5.5	Practical Deployment Considerations	105
3.5.6	Conclusion	106
4	Impact of Synthetic Image Generation on CNN Performance and Generalization	107
4.1	Introduction	108
4.1.1	Motivation for Synthetic Data in Cultural Heritage	108
4.1.2	Research Questions and Hypotheses	110
4.2	Synthetic Image Generation Framework	111
4.2.1	Two-Stage Generation Approach	111

4.2.2	LoRA Model Training	113
4.2.3	Synthetic Image Generation with LoRA	118
4.2.4	Quality Control and Filtering	121
4.3	Experimental Design	125
4.3.1	Synthetic Dataset Creation	125
4.3.2	Training Configurations	127
4.3.3	Evaluation Framework	128
4.4	Results and Analysis	129
4.4.1	Synthetic Image Quality Assessment (RQ3.1)	129
4.4.2	Impact on Class Imbalance (RQ3.2)	132
4.4.3	Mixed Training Effects (RQ3.3)	136
4.5	Comparative Analysis	139
4.5.1	Incremental Benefits of Synthetic Augmentation	139
4.5.2	Task-Specific Effectiveness	144
4.6	Discussion and Conclusion	150
4.6.1	Summary of Key Findings	150
4.6.2	Theoretical and Practical Implications	151
4.6.3	Limitations and Future Work	152
5	Conclusion	155
I	Appendix	160
A	Appendix A	161
	Bibliography	165

List of Figures

2.1	Literature selection process following the PRISMA methodology	14
2.2	Samples of typical Chinese porcelain ewers in ID3 [100]	19
2.3	Samples of archaeological porcelain pieces in ID8 [56]	19
2.4	Distribution of the dynasties in the porcelain datasets	20
2.5	Samples of TWW sherds in ID14 [79]	21
2.6	Samples of CMC Sukhothai ceramics dataset in ID12 [47]	22
2.7	Frequency of Image Processing Tasks in Reviewed Literature	26
2.8	Distribution of Feature Extraction Methods in Reviewed Papers	27
2.9	Distribution of Image Segmentation Methods in Reviewed Papers	28
2.10	Trend of CNN Model Methods in Feature Extraction and Image Segmentation	29
2.11	Distribution of Applications in Reviewed Papers	32
2.12	Tendency of ML/DL Algorithms on Image Classification	33
3.1	Map with the provenance of the porcelain Kilns Used in the Dataset	43
3.2	The Chinese porcelain glaze colour categories from the dataset	56
3.3	The Chinese porcelain form categories from the dataset	57
3.4	The Chinese porcelain pattern categories from the dataset	59
3.5	Hierarchical Label System for Porcelain Classification	60
3.6	Adaptive Dataset Splitting Strategy Based on Combination Size	62
3.7	Long-tail Distribution of Porcelain Combinations	63
3.8	Data Augmentation Effects on Representative Porcelain Samples	66
3.9	Architectural comparison between baseline and transfer learning approaches	68
3.10	Multi-Task Learning Framework Conceptual Diagram	70
3.11	Architecture Comparison: Key Building Blocks	73
3.12	Loss configuration and class distribution analysis	76
3.13	Comprehensive Performance Metrics Heatmap	84
3.14	Training Dynamics of Transfer Learning Models	85
3.15	MobileNetV3 Multi-Task Confusion Matrices on Test Set	86
3.16	MobileNetV3: Per-Class F1 Performance Analysis	88
3.17	Impact of Class Distribution on MobileNetV3 Performance	90
3.18	Top Classification Confusion Pairs by Task	93

3.19 Error Type Taxonomy Analysis	94
3.20 Error Rate vs Training Sample Size Analysis	96
3.21 Multi-Task Grad-CAM Analysis: Visualizing Model Attention in Porcelain Classification	97
3.22 A gradient-based feature importance visualization for the MobileNetV3 model on Sample 135, illustrating the regions of interest for (a) Dynasty classification (binary Song vs. Yuan), (b) Kiln attribution (17 production sites including Ding kiln), (c) Glaze classification (17 glaze types showing White and Ivory White confusion), and (d) Vessel type classification (20 morphological categories with Plate as primary prediction)	99
3.23 Attention Region Statistics in Multi-Task Porcelain Classification	100
4.1 LoRA training curves showing loss convergence and learning rate scheduling with three cosine restarts	117
4.2 Synthetic image generation pipeline with LoRA-enhanced Stable Diffusion	120
4.3 Dataset Composition Comparison	125
4.4 Visual Comparison: Real vs. Synthetic Porcelain Images	130
4.5 t-SNE Visualization of Real vs. Synthetic Porcelain Features	131
4.6 Class Distribution Before and After Augmentation	133
4.7 Lorenz Curve - Class Distribution Equality Before and After Synthetic Augmentation	134
4.8 Mixed Training Dynamics and Convergence Analysis	137
4.9 Training Stability (Loss Variance)	138
4.10 Task-wise Learning Dynamics and Convergence Patterns	139
4.11 Confusion Matrix Analysis - Comparing Baseline and Real+Syn-2500 Across All Tasks	140
4.12 Incremental Benefit Curve	142
4.13 Augmentation Strategy Decision Tree	143
4.14 Minority vs. Majority Class Performance Gains with Synthetic Augmentation	145

List of Tables

2.1	Inclusion and exclusion criteria	13
2.2	Critical assessment results of the reviewed studies	16
2.3	Information extracted from the papers included in the systematic literature review	16
2.4	Dataset of the reviewed studies	18
2.5	Image processing of the reviewed studies	25
2.6	ML algorithms of the reviewed studies	31
2.7	Accuracy of DL algorithms on image classification	34
2.8	Accuracy of ML algorithms on image classification	35
3.1	Image Sources and Licences for Museum Digital Collections	42
3.2	Distribution of Song and Yuan dynasty porcelain images by data source	42
3.3	Major Song-Dynasty Kiln Systems and Representative Sites	44
3.4	Characteristics of Song Ru Kiln with Example Photos	45
3.5	Characteristics of Song Ge Kiln with Example Photos	45
3.6	Comparative Analysis of Jun Kiln Flowerpots: Song Dynasty versus Yuan Dynasty	46
3.7	Comparative Analysis of Song Ding White Glaze Decoration Techniques and Motifs	48
3.8	Comparative Analysis of Song Cizhou Kiln Decoration Techniques and Characteristics	49
3.9	Comparative Analysis of Longquan Celadon Characteristics Across Song Dynasties	51
3.10	Decorative Techniques of Song Jizhou Kiln Black Glazed Porcelain . . .	52
3.11	Comparative Analysis of Yuan Jun Kiln Decoration Techniques	53
3.12	Characteristics of Yuan Cizhou Kiln Porcelain	54
3.13	Comparative Analysis of Yuan Dynasty Jingdezhen Porcelain	55
3.14	Hierarchical Label System	61
3.15	Dataset Distribution Statistics	63
3.16	Individual Task Statistics	64
3.17	Task Characteristics and Challenges	71

3.18 Specifications of Pre-trained Models Adapted for Multi-task Porcelain Classification	74
3.19 Fine-tuning configurations by model and component	75
3.20 Optimization settings for multi-task training	78
3.21 Evaluation metrics used for multi-task porcelain classification	80
3.22 Software stack used for model development and evaluation	82
3.23 Comprehensive Test Results on Test Set	83
4.1 LoRA Training Data Selection Strategy	115
4.2 LoRA Training Configuration	116
4.3 Hierarchical Prompt Structure with Real Examples	119
4.4 Dual-Dataset Synthetic Generation Strategy	120
4.5 Prompt Variation Strategies Derived from Training Data	122
4.6 Quality Metrics Comparison – Vanilla SD vs. LoRA-Enhanced SD	123
4.7 Generation Success by Target Category	124
4.8 Impact on Rare Category Representation	126
4.10 Data Augmentation Parameters	127
4.9 CNN Architecture and Training Parameters	128
4.11 Evaluation Metrics and Their Purpose	129
4.12 Comparison of Class Imbalance Metrics Between Chapter 3 and Chapter 4	134
4.13 Impact of Synthetic Augmentation on Extreme Rare Classes	135
4.14 Synthetic Data Volume Impact on Class Balance	135
4.15 Training Efficiency and Convergence Characteristics	136
4.16 Validation Performance and Training Stability	137
4.17 Epoch-wise Validation F1 Progression for Different Dataset Configurations	138
4.18 Cost-Benefit Analysis of Augmentation Stages	141
4.19 Complementary Characteristics of Traditional and Synthetic Augmentation (with Sources)	142
4.20 Task-Specific Test Performance Impact	144
4.21 Task-Specific Learning Dynamics with Syn-2500	145
4.22 Minority vs. Majority Class F1 Improvements	146
4.23 Visual Feature Dependencies and Synthetic Impact	147
4.24 Key Confusion Pattern Changes with Syn-2500	149
4.25 Top-5 Accuracy Analysis	149
A.1 Details of the reviewed studies	162
A.2 Dataset details of the reviewed studies	163

Chapter 1

Introduction

1.1 Research Background

Chinese porcelain holds an esteemed place in both cultural heritage and art history, renowned for its craftsmanship, aesthetic value, and historical importance [114]. Dating back to the Tang and Song dynasties, porcelain artifacts reflect the evolution of techniques, glazes, and decorative styles over centuries. Accurately classifying these artifacts into categories such as dynasty, glaze, ware, and type is crucial for historians, archaeologists, and collectors, as it aids in provenance research, forgery detection, and understanding historical trade routes [21]. Traditional classification methods rely heavily on expert knowledge and manual inspection, which are time-consuming and prone to subjectivity [16].

The application of Artificial Intelligence (AI) within the cultural heritage domain has emerged as a transformative tool, offering opportunities to augment traditional methods [54]. AI-driven analysis can enhance artifact identification, restoration, and preservation efforts by processing vast collections rapidly and uncovering patterns that may elude human experts [83, 27]. For instance, Deep Learning (DL) techniques have been employed to restore damaged frescoes [70] and authenticate paintings [18]. These successes demonstrate AI's potential to revolutionize how we engage with and protect cultural heritage, making it more accessible to researchers and the public alike. Machine Learning (ML) algorithms have emerged as powerful tools not only for automating and enhancing artifact classification but also from a digital sustainability perspective. By digitizing and automating the classification process, this approach reduces the need for repeated physical handling of fragile artifacts, minimizing the risk of damage and ensuring long-term preservation of cultural heritage [67]. Furthermore, creating digital repositories of classified artifacts facilitates global accessibility, enabling researchers and institutions to collaborate and study porcelain collections without the environmental and financial costs associated with physical transportation. Such advancements align with sustainable digitization initiatives, fostering a balance between technological innovation and the

responsible stewardship of cultural resources for future generations [74]. Convolutional Neural Networks (CNNs), in particular, have demonstrated remarkable performance in visual recognition tasks [45], making them well-suited for analyzing porcelain images. By leveraging pretrained models such as InceptionV3 [102], MobileNetV3 [93], ResNet50, ResNet101 [32] and EfficientNetB2 [104], researchers can benefit from transfer learning, utilizing knowledge from large scale datasets like ImageNet [15] to improve performance on smaller, domain-specific datasets. This approach reduces training time and mitigates the need for extensive labeled data, which is often scarce in cultural heritage studies.

Recent advancements in AI have opened new avenues for the intelligent identification of archaeological porcelains. Given the cultural significance of these artifacts, accurately identifying and classifying them is crucial for historical analysis and preservation. AI-driven methods for porcelain identification can be broadly categorized into ML-based and DL-based approaches. Due to the delicate and valuable nature of archaeological porcelains, these methods predominantly rely on digital images rather than physical handling of the artifacts. Traditional ML techniques have been widely employed for porcelain classification. Zhang et al. [119] tackled the problem of dynasty classification. They employed a SVM to classify porcelain images in Yuan, Ming and Qing dynasties, achieving respectively an accuracy of 75%, 80%, and 90%. Then, Zhou and Wang [127] proposed a novel semi-supervised clustering method to address the challenge of achieving accurate classification with individual image tiles. The technique is based on a Mean Shift Clustering algorithm that uses sample constraints to compute an optimal mapping kernel, ultimately resulting in an 74% of accuracy, a more precise classification compared to common fuzzy clustering methods. Finally, Hein et al. [34] explored the use of automated image analysis for petrographic characterization. The study utilized various classic texture features, such as those derived from Gabor filters, Laws' texture measures, and Haralick's texture measures, to classify porcelain materials from different Egyptian sources. The researchers achieved an accuracy of nearly 74% using a simple K-NN classifier. While ML approaches are computationally efficient, they rely heavily on hand-crafted feature extraction, which limits their ability to handle the intricate visual details of porcelain artifacts.

More recent studies have demonstrated the effectiveness of DL, particularly CNNs, in the feature extraction and classification of porcelain images. Yang et al. [116] employed a custom CNN for porcelain type recognition, achieving an accuracy of 70%. They compared the performance of their DL approach against a traditional ML approach (i.e., logistic regression). Also Ma et al. [61] focused on type recognition of ancient porcelains. They proposed a novel identification method called VGG16-GSSVM,

which is specifically designed for few-shot learning. This approach uses a pre-trained VGG16 model to extract shape features and then employs an SVM classifier, with its parameters optimized using a grid search algorithm. The experiment demonstrated that the VGG16-GSSVM model surpassed the accuracy and efficiency of traditional CNNs and CNN-SVM models, reaching an accuracy of 89%. Always in the context of porcelain type classification, Yue et al. [118] compared the performance of Inception, ResNet, and MobileNet, three well-known CNNs. The study found that all three models were highly effective at recognizing and classifying porcelain types, achieving an accuracy of up to 90%.

Despite the promising results achieved by prior studies, dataset limitations remain a significant bottleneck in archaeological porcelain classification. Weng et al. [110] demonstrated high accuracy using a dataset of only 50 images, highlighting the small scale of many existing datasets. Such datasets often lead to overfitting, reduced generalization, and poor performance on unseen data. Moreover, many studies restrict feature selection to narrow characteristics. For example, Weng et al. [110] focused on porcelain of the same shape from different dynasties, while Zhang et al. [119] examined artifacts with similar patterns. This reductionist approach overlooks the rich diversity of Chinese archaeological porcelain, which varies significantly in glaze, texture, composition, and form due to differences in raw materials, production methods, and firing techniques [112]. To achieve more comprehensive and reliable classification, future datasets must incorporate diverse features representative of the full spectrum of porcelain artifacts, fostering robust and generalizable model performance.

1.2 Research Questions

Within the intersection of artificial intelligence and cultural heritage preservation, this thesis explores how machine learning, particularly deep learning techniques, can revolutionize the identification and classification of Chinese archaeological porcelain. This study aims to bridge the gap between traditional expert-based methods and modern computational approaches, while addressing the unique challenges posed by limited datasets and the multi-faceted nature of porcelain classification.

RQ-1: *How can machine learning methods be effectively applied to the identification of Chinese archaeological porcelain, given the current limitations of available data?*

The application of machine learning to archaeological ceramics has shown promising yet varied results across different approaches. While traditional ML methods have achieved moderate success rates between 70-90% [119], they remain constrained by

their reliance on hand-crafted features. This research question systematically examines the landscape of ML applications in ceramic identification, seeking to understand not just what has been achieved, but what methodological patterns emerge from successful implementations. By analyzing the progression from classical ML to deep learning approaches, we can identify the critical factors that determine classification success and the persistent challenges that limit current methods.

RQ1.1: *What is the state of the art in machine-learning-based identification of archaeological ceramics, and what methodological or data-related gaps remain?*

Dataset limitations represent perhaps the most significant bottleneck in archaeological porcelain classification. Current studies often work with extremely limited datasets, some containing as few as 50 images total [110], which severely constrains model generalization. This sub-question investigates the characteristics, scale, and diversity of existing datasets, examining how their limitations influence classification outcomes. Understanding the current dataset landscape is crucial for developing strategies to overcome data scarcity through techniques like transfer learning and synthetic data generation.

RQ1.2: *What types of datasets, image-processing techniques, and learning algorithms have proven most effective in previous studies, and how do they inform the design of a robust porcelain-classification framework?*

The visual complexity of porcelain artefacts encompassing variations in glaze, texture, decorative patterns, and form presents unique challenges for image analysis. Different preprocessing techniques, feature extraction methods, and image enhancement strategies can significantly impact classification accuracy. This question explores how various image processing approaches handle the intricate visual details of porcelain [39], from traditional texture analysis using Gabor filters and Haralick features to modern CNN-based automatic feature learning.

RQ1.3: *What are the main ML algorithms used in the identification of archaeological ceramics?*

From Support Vector Machines achieving 75-90% accuracy in dynasty classification [119] to custom CNNs reaching 94% accuracy in kiln recognition [55], the choice of algorithm profoundly impacts classification performance. This sub-question systematically evaluates the spectrum of ML algorithms applied to ceramic identification, comparing their strengths, limitations, and suitability for different classification tasks. Special attention is given to the evolution from traditional ML to deep learning approaches and the trade-offs between computational efficiency and classification accuracy.

RQ-2: *To what extent can deep learning models (especially CNN-based multi-task*

architectures) accurately classify key attributes of Chinese porcelain under real data constraints?

Unlike previous studies that focus on single classification tasks, real-world porcelain identification requires simultaneous consideration of multiple attributes. A single artifact must be classified by dynasty, glaze type, ware, and specific type, each attribute providing crucial information for historical analysis. This research question investigates whether CNNs can effectively handle this multi-dimensional classification challenge through multi-task learning, where shared representations across related tasks could lead to improved data efficiency and better generalization. The focus on Song and Yuan dynasties specifically addresses periods of significant technological and artistic innovation in Chinese porcelain production.

RQ2.1: *What performance can be achieved when simultaneously classifying dynasty, kiln provenance, glaze, and vessel form?*

While individual studies have reported accuracies ranging from 70% to 94% for specific classification tasks [55], no comprehensive evaluation exists for simultaneous multi-attribute classification. This sub-question establishes baseline performance metrics for deep learning models across all four classification dimensions, providing crucial benchmarks for evaluating the effectiveness of various architectural choices and training strategies.

RQ2.2: *How does transfer learning influence classification performance under limited, imbalanced archaeological datasets?*

Transfer learning has emerged as a critical strategy for addressing data scarcity in cultural heritage applications. By leveraging knowledge from models pre-trained on ImageNet, researchers can potentially overcome the limitations of small archaeological datasets. This question quantifies the performance gains achieved through transfer learning across different CNN architectures (ResNet50, ResNet101, MobileNetV3, EfficientNetB2, and InceptionV3), comparing feature extraction versus fine-tuning approaches and determining optimal strategies for each classification task.

RQ2.3: *How can explainable AI methods help interpret model decisions and relate computational features to archaeologically meaningful attributes?*

The "black box" nature of deep learning models poses a significant barrier to their adoption in cultural heritage research, where understanding the reasoning behind classifications is as important as accuracy. Art historians and archaeologists need to know which visual features drive classification decisions to validate results against expert knowledge and potentially discover new patterns. This question explores how explainable AI techniques can bridge the gap between computational efficiency and

human interpretability, making AI tools more trustworthy and useful for domain experts.

RQ-3: *Can synthetic image generation mitigate data scarcity and imbalance in archaeological porcelain datasets without compromising classification reliability?*

The scarcity of labeled porcelain images severely limits the potential of deep learning approaches. Synthetic data generation offers a promising solution, but its effectiveness in capturing the subtle visual characteristics of authentic porcelain remains unexplored. This research question investigates whether artificially generated porcelain images can augment limited real datasets without introducing artifacts or biases that compromise classification accuracy. The answer has implications not only for porcelain classification but for the broader application of synthetic data in cultural heritage preservation.

RQ3.1: *How do the quality and diversity of generated images affect CNN training?*

Not all synthetic images are created equal. The fidelity with which generated images capture authentic porcelain characteristics, including realistic glaze effects, appropriate wear patterns, and accurate decorative motifs, directly impacts their utility as training data. This sub-question evaluates different generation techniques and quality metrics, establishing guidelines for creating synthetic datasets that genuinely enhance model performance.

RQ3.2: *Can synthetic images effectively balance underrepresented classes and improve performance on rare kiln, glaze, or form categories?*

Archaeological datasets often suffer from severe class imbalance, with certain rare types or dynasties poorly represented. Synthetic data generation could theoretically address this imbalance by generating additional samples for underrepresented classes. This question quantifies the effectiveness of this approach, examining whether synthetic augmentation can improve classification performance for rare categories without degrading overall accuracy.

RQ3.3: *What is the optimal combination of real and synthetic data to maximise model stability and generalisation?*

The optimal balance between real and synthetic data remains an open question. Too much synthetic data might cause models to learn artificial patterns, while too little may not address data scarcity effectively. This sub-question systematically investigates different mixing ratios, evaluating their impact on model robustness, training stability, and ability to generalize to unseen authentic porcelain images. The findings will provide practical guidelines for incorporating synthetic data into training pipelines while maintaining classification reliability.

1.3 Outline

Given the background concepts exposed above, a detailed overview of the work presented in this thesis is provided below:

Chapter 1 The first chapter introduces the context of the thesis and presents the research questions. Lastly, it summarizes the content of the thesis.

Chapter 2 This chapter presents our systematic literature review on machine learning for archaeological ceramics identification, encompassing both pottery and porcelain. While the review covers the broader ceramic domain to provide comprehensive context, this thesis specifically focuses on porcelain classification. The findings related to porcelain (16 out of 33 studies) directly inform our research, while pottery studies provide valuable methodological insights applicable to our work. This chapter aims to answer the research question **RQ-1**, **RQ-1.1**, **RQ-1.2** and **RQ-1.3**.

Chapter 3 This chapter explores the performance of these CNN architectures in classifying porcelain artifacts across multiple tasks: dynasty, glaze, kiln, and type. It presents a comparative analysis of models, evaluating their performance. Furthermore, the impact of transfer learning is investigated by comparing models trained with and without pre-trained weights. Additionally, we incorporated eXplainable Artificial Intelligence (XAI) techniques to move beyond simple performance metrics and provide valuable insights into the models' decision-making processes by visualizing and quantifying the features they rely on for classification. The goal of the chapter is to provide an answer to the research question **RQ-2**, **RQ-2.1**, **RQ-2.2** and **RQ-2.3**.

Chapter 4 This chapter investigates the potential of synthetic image generation to enhance CNN-based porcelain classification through three interconnected research questions. Through this investigation, we aim to establish whether synthetic data generation can effectively address the fundamental challenges identified in Chapter 3 while maintaining the archaeological integrity essential for cultural heritage applications. This chapter aims to answer the research question **RQ-3**, **RQ-3.1**, **RQ-3.2** and **RQ-3.3**.

Chapter 5 This chapter ends the thesis by drawing conclusions and paves the way for future contributions.

Chapter 2

Machine Learning for Porcelain Classification: A Systematic Review

How can machine learning methods be effectively applied to the identification of Chinese archaeological porcelain, given the current limitations of available data?

What is the state of the art in machine-learning-based identification of archaeological ceramics, and what methodological or data-related gaps remain?

What types of datasets, image-processing techniques, and learning algorithms have proven most effective in previous studies, and how do they inform the design of a robust porcelain-classification framework?

What are the main ML algorithms used in the identification of archaeological ceramics?

— RQ-1, RQ-1.1, RQ-1.2, RQ-1.3

This chapter presents our systematic literature review on machine learning for archaeological ceramics identification, encompassing both pottery and porcelain. While the review covers the broader ceramic domain to provide comprehensive context, this thesis specifically focuses on porcelain classification. The findings related to porcelain (16 out of 33 studies) directly inform our research, while pottery studies provide valuable methodological insights applicable to our work.

2.1 Introduction and Background

Ceramics refer to a broad category of materials, typically inorganic and non-metallic, employed by humans for over 10,000 years [88]. It includes pottery, majolica, faience, terracotta, stone mass, and porcelain [25]. Their remarkable durability has allowed them to withstand the test of centuries, preserving a window into the past and safeguarding the stories of our ancestors [9]. Consequently, ceramics serve as chronological markers that are indispensable for reconstructing the past and comprehending the temporal progression of cultures, from their emergence to their decline, thereby enriching our understanding of the human journey through time [86]. Moreover, since they were among the earliest commodities traded between different regions and cultures, they serve as invaluable artefacts that unveil intricate trade networks and interactions between ancient societies [68]. They trace the paths of cultural exchange, uncover the complexities of ancient trade routes, and highlight the interconnections of diverse civilizations across time and space [9]. For these reasons, it becomes of fundamental importance to accurately date and classify ancient ceramics.

Currently, the methods for the identification of archaeological ceramics, mainly focused on pottery and porcelain, can be grouped into two main families [107]. On the one hand, there is empirical identification [60]. It is a skilled process relying on the expertise of human analysts, who identify archaeological ceramics manually. Hence, it is a time-consuming activity that often lacks objective constraints, leading to inconsistent classifications. On the other hand, there are scientific identification methods that encompass techniques such as X-ray fluorescence analysis, thermoluminescence dating, and spectral analysis [111]. Thermoluminescence dating may cause damage to the porcelains themselves, resulting in irreversible damage [81]. While X-ray fluorescence analysis is a non-destructive technique [6], it does have some drawbacks for definitive identification: i) it primarily detects elements present in the porcelain and it misses subtle details, including firing temperature, specific minerals used, or manufacturing techniques; ii) it hinges on the quality and completeness of the reference database used about the elemental makeup of known porcelains; iii) it typically analyzes the surface composition that can be altered because of weathering, restoration, or glazing with a different composition. Moreover, it requires expensive equipment, involves complex identification processes, and is time-consuming.

Hence, the identification of archaeological ceramics, including pottery and porcelain, faces two key challenges: it can be time-consuming and it heavily relies on expert knowledge and experience. This has led researchers to explore Computer Vision (CV)

and Artificial Intelligence (AI) as potential tools to aid in the identification process. Thanks to the advancements in these fields [23, 2, 53], more studies investigated their use for the identification of archaeological ceramic images in recent years [105, 99, 1]. Just to cite a few, Wang et al. [109] applied the Support Vector Machine (SVM) algorithm to design a prototype for feature extraction and classification of porcelain images. Instead, Zheng et al. [125] employed a Gaussian colour model and multi-scale filter bank to extract colour texture features.

The application of ML and DL techniques in the field of archaeological ceramics identification is an interesting area of research. While our review seeks to provide a comprehensive overview, it is important to acknowledge previous works that have explored similar themes. Notably, several surveys and studies have been conducted to examine the intersection of archaeological analysis and computational techniques. These existing reviews serve as foundational references, offering valuable insights and highlighting the progress made in the application of ML and DL methodologies to archaeological ceramics identification.

Cumbajin et al. [14] presented a systematic literature review on deep learning for industrial surface defect detection of several materials, including ceramics. Such a review considered only the use of convolutional neural networks. Our paper differentiates itself by focusing exclusively on the identification and classification of archaeological ceramics, employing a variety of ML and DL algorithms tailored to the unique characteristics of archaeological data. While the industrial review proposes a new taxonomy for surface defect detection, our work seeks to develop and refine methodologies specifically for archaeological ceramics.

Zhang et al. [120] proposed a review on the use of ML the discovery of high-entropy ceramics. High-entropy materials, known for their versatile platform and superior structural and functional properties, present a vast phase space that complicates efficient design through traditional methods. The review highlights how ML accelerates the discovery and optimization of high-entropy ceramics. The paper discusses the entire ML pipeline from data collection and feature engineering to model refinement and performance prediction improvement. In contrast, our review delves into the application of ML algorithms for archaeological ceramics, which, unlike high-entropy ceramics, involves the analysis of artifacts with historical and cultural significance.

Di Angelo, Di Stefano, and Guardiani in [16] conducted a comprehensive review of automatic methods for analyzing archaeological pottery sherds, encompassing the state-of-the-art up to the end of 2021. They provided a critical analysis of the most significant advancements in pottery analysis, classification, and reconstruction from a

3D discrete manifold model. In contrast to this review, our paper do not focus on 3D models but mainly on images.

Yue et al. [71] provided a review of the application of virtual reality (VR) technology in the conservation and restoration of archaeological ceramics. However, this review does not cover the application of ML algorithms for the identification of archaeological ceramics, although the existing review on VR technology is not directly related to ML, it still can provide future research directions in the field of archaeological ceramics, such as integration of ML algorithms with VR for the identification of archaeological ceramics.

This chapter presents a systematic literature review on machine learning for archaeological ceramics identification, structured as follows: Section 2.2 describes the PRISMA methodology and how the data were collected and processed. Section 2.3 outlines the main datasets employed in the identification of archaeological ceramics. Section 2.4 presents the findings about the image processing of archaeological ceramics for its identification. Section 2.5 provides the main ML algorithms used in the identification of archaeological ceramics. Section 2.6 discusses this chapter by concluding.

2.2 Methodology

This section presents the research questions that motivated this research, describes the PRISMA methodology and how the data were collected and processed.

2.2.1 PRISMA Methodology

Systematic reviews aim to identify all research addressing a specific and well-defined question to provide a balanced and unbiased literature summary [75]. A systematic review takes advantage of a rigorous and transparent methodology, often outlined in a protocol, to identify, evaluate, and synthesize relevant research studies. The search for relevant studies is often carried out across various databases and sources. Finally, the studies are evaluated and the main findings are synthesized to provide a comprehensive overview.

In this work, we took advantage of the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) methodology [72, 77]. PRISMA provides a framework for the conduct of systematic reviews and future analyses in the field of research. It ensures transparency, reproducibility and decrease of bias in the review process and including an organized and complete approach to the organisation and presenting of systematic reviews.

2.2.2 Data Collection and Processing

We checked published English-language articles or materials that are available on the web through noticeable digital libraries and databases in the field of computer science. They include ACM Digital Library, IEEE Xplore, Scopus, ScienceDirect, SpringerLink, and CNKI. In addition to them, Google Scholar was used to supplement the search to minimise the risk of exclusion.

The following search keywords (Boolean logic) were entered into the six databases: (“computer vision” OR “artificial intelligence” OR “machine learning” OR “deep learning”) AND (“archaeological ceramic” OR “ancient ceramic” OR “porcelain”) AND (“identification” OR “classification” OR “recognition”). The search options were title, abstract and keywords for the Scopus database.

We included any conference or journal paper concerning computer vision and machine learning for the identification of archaeological ceramics that have been published during the last ten years (i.e., between 2013 and 2023). To ensure both quality and accuracy, only peer-reviewed journal articles for which the full text is available have been included. The other inclusion and exclusion criteria are listed in Table 2.1. It is important to notice that ML algorithms and general computer vision approaches must be employed to identify or classify archaeological ceramics and not for other tasks such as reconstruction or restoration. Moreover, we are interested in finding research articles and not other literature reviews.

Table 2.1. Inclusion and exclusion criteria

Inclusion Criteria	Exclusion Criteria
Focus on image processing techniques and feature extraction approaches for archaeological ceramics images	Focus on chemical elements or spectra of archaeological ceramics
Use computer vision and machine learning approaches for identification or classification of archaeological ceramics	Use computer vision and machine learning approaches for reconstruction or restoration of archaeological ceramics
Empirical studies	Literature reviews, commentaries or meta-analysis
Written in English	Written in other languages
Peer-review	Pre-print

The papers were searched by following four steps: i) collection, ii) screening, iii) eligibility, and iv) information extraction. A summary of the papers selected and excluded at each step of the PRISMA review process is reported in Figure 2.1. Firstly,

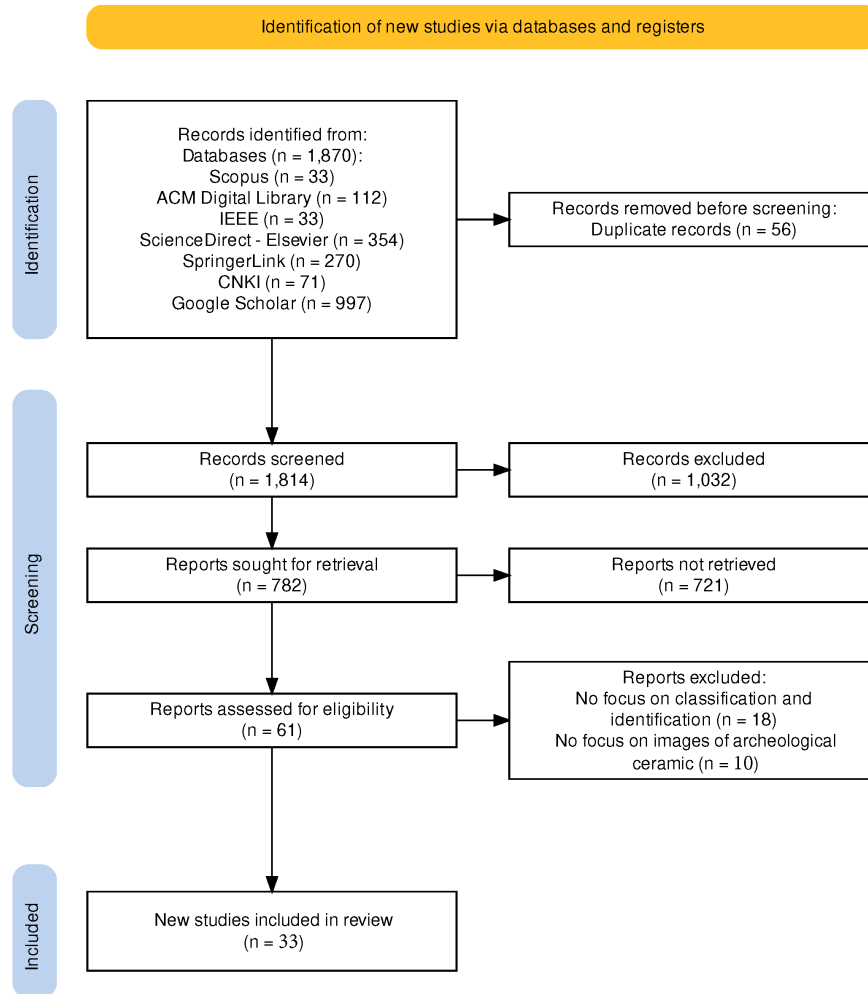


Figure 2.1. Literature selection process following the PRISMA methodology

we collected all the papers that corresponded to the search criteria. Secondly, the dataset with the most results was Google Scholar with 997 papers. The duplicated records removed (i.e., 56 papers) were found in Scopus, ACM Digital Library, IEEE Explore, Science Direct, Springer Link, and CNKI, and were also found on Google Scholar. From the remaining 1814 papers, we excluded the following ones: 1032 papers because i) the language of the paper is not English, and ii) the title indicates that the paper is not focused on the identification of archaeological ceramics using computer vision techniques. 721 papers because at the abstract level, they were not focused on archaeological ceramics but modern ceramics. 18 papers because they were not focused on identification and classification, but on other tasks. 10 papers because they were not focused on images of archaeological ceramics. Finally, thirty-three papers meeting

the standard were eventually retained (listed in Table A.1 in Appendix A), labelled ID1-ID33 sequentially.

2.2.3 Critical Assessment

Besides the selection process described in the subsection 2.2.2, an assessment of the quality criteria was also performed. The results are presented in Table 2.2, which has one column for each research question (RQ1.1, RQ1.2, and RQ1.3) and as many rows as articles. For each article resulting from the selection process, we assessed whether they included an answer to the different research questions (number 1 in the table) or not (number 0 in the table). Regarding RQ1, which focuses on the dataset, most of the papers well described the data details of the study, while three studies, namely ID5 [26], ID10 [48], and ID17 [33], did not provide adequate information. With regard to RQ2, out of thirty-three papers, only ID13 [116] did not evidently describe the image processing methods used. Regarding RQ3, five studies lacked a clear description of the ML algorithm for the identification of ceramics. In conclusion, twenty-six studies provide enough information to answer all the research questions while the minority of studies (i.e., five ones) obtain a score of two out of three, and two studies ID10 [48] and ID17 [33] obtain a score of one out of three.

2.2.4 Extraction of Relevant Fields

Finally, the last step of our methodology involves eliciting and summarising relevant information from the selected articles. This was done manually by reading the full texts of the selected articles. The elicited information includes the answers to the research questions of the article. The complete list of fields elicited from the articles is provided in Table 2.3. Three fields, including RQ1.1, RQ1.2 and RQ1.3, were manually compiled in response to the different research questions.

2.3 Datasets for Porcelain Identification(RQ1.1)

2.3.1 Datasets for archaeological ceramics identification

This Section answers the first research question **RQ1.1**, "What is the state of the art in machine-learning-based identification of archaeological ceramics, and what methodological or data-related gaps remain?" The answer includes six main aspects of the selected

Table 2.2. Critical assessment results of the reviewed studies

ID	Reference	RQ1.1	RQ1.2	RQ1.3	Total
ID1	[55]	1	1	1	3
ID2	[69]	1	1	1	3
ID3	[100]	1	1	1	3
ID4	[118]	1	1	1	3
ID5	[26]	0	1	1	2
ID6	[30]	1	1	1	3
ID7	[76]	1	1	0	2
ID8	[56]	1	1	1	3
ID9	[10]	1	1	1	3
ID10	[48]	0	1	0	1
ID11	[19]	1	1	1	3
ID12	[47]	1	1	1	3
ID13	[116]	1	0	1	2
ID14	[79]	1	1	1	3
ID15	[117]	1	1	1	3
ID16	[61]	1	1	1	3
ID17	[33]	0	1	0	1
ID18	[60]	1	1	1	3
ID19	[17]	1	1	1	3
ID20	[119]	1	1	1	3
ID21	[73]	1	1	0	2
ID22	[126]	1	1	1	3
ID23	[34]	1	1	1	3
ID24	[8]	1	1	1	3
ID25	[44]	1	1	1	3
ID26	[63]	1	1	1	3
ID27	[127]	1	1	1	3
ID28	[110]	1	1	0	2
ID29	[85]	1	1	1	3
ID30	[105]	1	1	1	3
ID31	[80]	1	1	1	3
ID32	[3]	1	1	1	3
ID33	[65]	1	1	1	3

Table 2.3. Information extracted from the papers included in the systematic literature review

	Field	Description
1	ID	Unique identifier for the study
2	Title	Title of the paper
3	Authors	Authors of the paper
4	Year	Year of publication
5	Country	Country of the first author of the paper
6	Keywords	The keywords listed by the authors in the paper
7	Main results	Main results presented in the paper
8	Information about the quality analysis	Criteria presented in subsection 2.2.3
9	RQ1.1	Datasets employed in the archaeological ceramics identification for computer vision
10	RQ1.2	Findings related to the image processing of archaeological ceramics for identification
11	RQ1.3	ML algorithms used in the identification of archaeological ceramics

dataset: ceramic type (porcelain and pottery), country, period, pattern or type, sample count, and data source. Table 2.4 presents the dataset employed in each paper. As expected, the mass of them (30 out of 33 ones, that is 90.9%) clearly designated the dataset used. For the other three papers (9.0%), ID5, ID10, and ID17, the label "N/A" is used, since they were mainly focused on the image processing methods [26, 48].

As already anticipated, in the field of archaeology, ceramics include pottery, majolica, faience, terracotta, stone mass, and porcelain [25]. In the selected studies, the datasets mainly contain two categories, pottery and porcelain, which are distinguished in Table 2.4 by the labels "Pottery" and "Porcelain". If the type of ceramic is not specified, we used the label "Ceramic". For the other characteristics (Period, Pattern Types, Sample Count, Data Source), they are set to N/A if in the original section, such information is not specified. More detailed information about the dataset is listed in Table A.2 in the Appendix A.

2.3.2 Type, country, and period of ceramics

These datasets mainly contain two different categories of ceramics, porcelain and pottery. They cover a wide range of periods (from the Neolithic period to the early 19th century). The datasets belong to 12 countries, including the United States of America, China, France, Greece, Egypt, Honduras, Italy, Japan, Korea, New Zealand, South Africa, and Thailand.

Porcelain is a special type of ceramic that originated in China during the Tang dynasty (ca.1332 to 1043 aBP) [25]. Among thirty papers, sixteen papers selected Chinese porcelains as the dataset, ranging from Tang to Song (ca.990 to 671 aBP), Yuan (ca.679 to 582 aBP), Ming (ca.582 to 306 aBP), and Qing (ca.306 to 38 aBP) dynasty. Figure 2.4 shows the distribution of dynasties in the porcelain datasets. Two papers focused on only one dynasty for the identification of porcelains. ID1 [55] classified the porcelains from the Yaozhou kiln of the Song dynasty while ID21 [73] identified the washer-type porcelains from the Ming dynasty. Meanwhile, seven papers discussed two or more dynasties for the identification of porcelains. ID3 [100] discussed the porcelain ewers from the Tang, Song, and Yuan dynasties, Figure 2.2 shows the samples of dataset in ID3. ID4 [118] classified porcelains with four different patterns from Ming and Qing dynasties, ID8 [56] identified porcelain shards from Yuan, Ming, and Qing dynasties, Figure 2.3 shows the samples of dataset of ID8, and ID11 [19] classified two types of porcelain vases from Ming and Qing dynasties. Moreover, ID13 [116] identified different types of porcelain from Song, Yuan, Ming, and Qing dynasties, ID20 [119] classified porcelains with Banana leaf patterns from Yuan, Ming, and Qing dynasties, and ID28

Table 2.4. Dataset of the reviewed studies

ID	Type	Country	Period	Pattern Types	Sample Count	Data Source
ID1 [55]	Porcelain	China	Song dynasty	N/A	N/A	Museum collection
ID2 [69]	Pottery	Japan	Late jomon period	N/A	1036 X-ray images	Not provided
ID3 [100]	Porcelain	China	Tang, Song, and Yuan dynasties	Ewers	232 images	Museum collection
ID4 [118]	Porcelain	China	N/A	19 types of vases	N/A	Baidu PaddlePaddle
ID5 [26]	Porcelain	China	N/A	N/A	N/A	N/A
ID6 [30]	Porcelain	China	Ming and Qing dynasties	Dragon, Fish, Peony, Lotus patterns	N/A	Museum Website
ID7 [76]	Porcelain	China	N/A	Various	5834 images	Online Repository
ID8 [56]	Porcelain	China	Yuan, Ming, and Qing dynasties	Shards by incomplete patterns	373 images	Scholarly book Jingdezhen relics
ID9 [10]	Pottery	Italy	Between the Late bronze age and the Orientalizing period	Vessel profiles	4992 images	Latium Vetus and Etruria archaeological sites
ID10 [48]	Porcelain	China	N/A	N/A	N/A	N/A
ID11 [19]	Porcelain	China	Ming and Qing dynasties	Mei vase, Yuhuchun vase	248 images	Beijing Museum and Taipei Museum website
ID12 [47]	Ceramic	Thailand	N/A	7 types of patterns	557 images	Museum collection
ID13 [116]	Porcelain	China	Song, Yuan, Ming, and Qing dynasties	Blue and white porcelain, Purple sand, Doucai	2750 images	Palace Museum website
ID14 [79]	Pottery	USA	ca.1125 to 650 aBP	Tusayan White Ware sherds	3064 images	Museum collection
ID15 [117]	Ceramic	Korea	N/A	Various	19610 images	Museum collection
ID16 [61]	Porcelain	China	N/A	7 types of bottles	5624 images	Baidu crawled data
ID17 [33]	Porcelain	China	N/A	N/A	N/A	N/A
ID18 [60]	Pottery	Honduras	Cocal period	Various	1069 images from 66 pottery sherds	Guadalupe archaeological site
ID19 [17]	Pottery	France	Early Neolithic period	Coiling and Spiral Patchwork	16 sherds	Pendimoun rock shelter
ID20 [119]	Porcelain	China	Yuan, Ming, and Qing dynasties	Banana leaf patterns	230 pieces	Museum collection
ID21 [73]	Porcelain	China	Ming dynasty	Washer	N/A	Academic research
ID22 [126]	Pottery	USA	Pre-Colonial	Paddle impressions on pottery sherds	N/A	North America archaeological sites
ID23 [34]	Pottery	Egypt	Nile clay and Marl clay	Nile clay and Marl clay	69 samples	Museum Collection
ID24 [8]	Ceramic	New Zealand	N/A	Various	N/A	New Zealand sites
ID25 [44]	Pottery	South Africa	N/A	Zulu pottery fabrics	50 per class	Field collection.
ID26 [63]	Ceramic	USA	ca.175 to 125 aBP	N/A	8 classes of fragments	Independence National Historical Park
ID27 [127]	Porcelain	China	N/A	N/A	288 pieces	Academic research
ID28 [110]	Porcelain	China	Tang, Song dynasty, Yuan, and Ming dynasties	Jars	50 images	Forbidden City Museum
ID29 [85]	Ceramic	USA	ca.175 to 125 aBP	N/A	6 pieces of fragments	Independence National Historical Park
ID30 [105]	Pottery	France	ca.1524 to 924 aBP	Wheel-carved	957 patterns	Saran archaeological site
ID31 [80]	Pottery	Greece	N/A	N/A	203 pottery sherds	Koroneia archaeological site
ID32 [3]	Ceramic	USA	ca.175 to 125 aBP	N/A	N/A	Independence National Historical Park
ID33 [65]	Pottery	Greece	Byzantine era	N/A	110 pottery sherd images	Byzantine antiquities

discussed porcelain jars from Tang, Song, Yuan, and Ming dynasties.

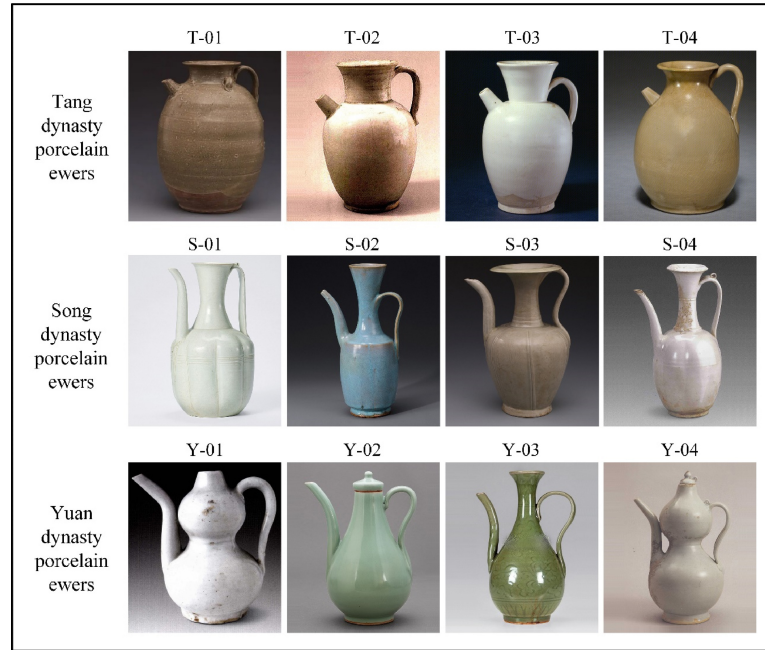


Figure 2.2. Samples of typical Chinese porcelain ewers in ID3 [100]



Figure 2.3. Samples of archaeological porcelain pieces in ID8 [56]

Different dynasties have different styles, and of course, the style of the porcelain depends on the kiln, so the shape of the porcelain, the colour, and its pattern will have different characteristics depending on the style. The Ming dynasty is the most studied dynasty, with seven papers focused on it. There are two main reasons for this choice. First, such a dynasty has the largest number of preserved porcelain samples. Second, it

represents the second flourishing period in the development of Chinese porcelain, which is valuable for research. Meanwhile, the Tang, Song, and Yuan dynasties had not been studied as much in comparison, which also provides a direction for future research. The direction can be expanded by focusing on the three less studied dynasties to broaden the scope of research. In addition, almost all of the selected papers focus on only one feature of porcelain, such as different types of porcelain from the same dynasty [19], or different patterns of porcelain from the same dynasty [30], or the study of the same type of porcelain from different dynasties [100]. No research integrates these features, such as different types of porcelain from different dynasties with different patterns, to expand the depth of the study of porcelain, which is also one of the directions that can be researched in the future.

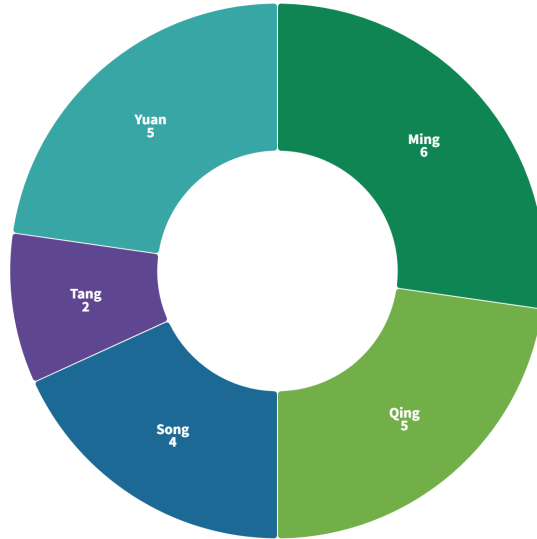


Figure 2.4. Distribution of the dynasties in the porcelain datasets

Pottery is another type of ceramics, eleven papers indicate that their chosen datasets were potteries, from Japan, Italy, the United States of America, Honduras, France, Egypt, South Africa, and Greece, ranging from the Early Neolithic period (ca. 7850 to 7450 aBP) to the early 19th century (ca. 175 to 125 aBP). Although from different countries, each paper selected only pottery samples from a single period to be studied and classified. One case is under the same period for pottery with different patterns, such as ID19 [17], which investigated potteries with coiling and spiral patchwork, ID14 [79] that focused on TWW pottery sherds with decorative patterns, Figure 2.5 shows the samples of dataset in ID14. The other case regards the same period but focuses on different types of pottery, such as ID9 [10] where the authors classified potteries of

different vessel profiles. At the same time, most of the datasets are focused on pottery sherds, which is one of the dilemmas faced by archaeological pottery. The manual classification of a large number of sherds is a time-consuming activity [60]. Hence, the main objective of the researchers is to reduce the time of pottery classification and improve the consistency of results.

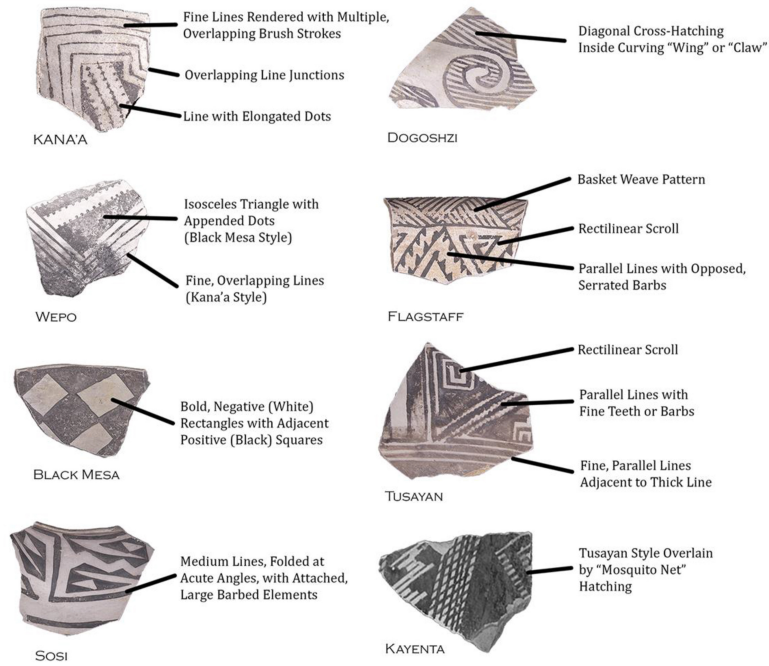


Figure 2.5. Samples of TWW sherds in ID14 [79]

The remaining six papers do not specify the type of ceramics or the period covered by their datasets. These studies originate from Thailand, Korea, New Zealand, and the United States of America, and primarily focus on ceramic patterns. For example, ID12 [47] classifies ceramics from the Sukhothai kiln into seven types of patterns. Figure 2.6 presents sample images from the dataset used in ID12.



Figure 2.6. Samples of CMC Sukhothai ceramics dataset in ID12 [47]

2.3.3 Data source and sample count of ceramic

The size of the dataset ranges from tens of samples to thousands of samples, depending on the preservation of the ceramics. For example, in ID2 [69], the number of samples preserved from this period was insufficient and therefore, the labelled data was insufficient, and self-built experimental data had to be used for the study. ID20 [119] has a similar situation with a limited number of available porcelain with banana leaf patterns. Therefore, the dataset size was only 230 samples. Having larger datasets enables the use of powerful algorithms, such as deep learning models, which typically yield superior results. The richness and volume of data allow these algorithms to learn more complex patterns and representations, thereby improving their accuracy and generalization capabilities.

As shown in Table 2.4 (column Data Source), the majority of the data were collected from the museum collection, which includes both images photographed in the museum

and images downloaded from the museum’s website, as well as datasets obtained from archaeological excavation sites. Additionally, a small number of datasets were obtained through web crawling (ID4, ID7 and ID16). The diverse methods of data collection ensure a broad and varied dataset, enriching the training material for the models. However, the quality of these datasets, as well as the trustworthiness of the labels of the images, is crucial. For instance, images obtained from official museum sources are likely to have accurate and reliable labels due to the expertise and meticulous documentation practices of museum professionals. Conversely, images from non-official museum websites or those collected via web crawling might have less reliable labels. This variability in label accuracy can significantly impact the training of machine learning models. Accurate labels are essential for the models to learn correctly, as they guide the algorithms in understanding and categorizing the images. Inaccurate labels can lead to incorrect learning, reducing the effectiveness and accuracy of the trained models. Therefore, ensuring the reliability of the labels, especially in self-built datasets, is vital for developing robust and effective machine learning models.

2.3.4 Data Scarcity and Future Directions

While ML algorithms have emerged as a promising tool for archaeologists, offering the potential to streamline the identification of archaeological ceramics, they still have to face a significant hurdle, data scarcity. In fact, to be truly effective, machine learning algorithms require a robust dataset of well-documented ceramic artefacts. This data should include detailed information on the origin, type, composition, and visual characteristics of each ceramic piece. Unfortunately, such comprehensive datasets are often scarce in archaeology.

Several factors contribute to data scarcity. On the one hand, archaeological excavations are often resource-intensive, and not all recovered ceramics receive in-depth analysis due to time and budget constraints. However, even if museums possess this data, several challenges remain. In fact, this information often exists in data silos, not readily transformed into usable datasets for training ML algorithms. Cataloguing and documenting ceramics can be a time-consuming and resource-intensive process. Many museums lack the staff or funding to create comprehensive, digital datasets for their entire collections. Finally, even if museums have digitized data, they might not readily share it for free due to concerns about ownership and potential commercial use.

To partially address this challenge, there are several initiatives to establish standardized recording practices for archaeological ceramics, with the aim of building and sharing open-access databases of ceramic data. Finally, the recent advancements in

multimodal Large Language Models able to generate images from text inputs could be exploited to generate synthetic data to augment the size of the datasets [42].

2.4 Image Processing Methods for Porcelain (RQ1.2)

This section aims to answer the second research question **RQ1.2**, "What types of datasets, image-processing techniques, and learning algorithms have proven most effective in previous studies, and how do they inform the design of a robust porcelain-classification framework?" Image processing is the process of transforming an image into a digital form and performing certain operations to get some useful information from it [39]. It encompasses a variety of techniques applied to digital images, typically in a sequential order, such as image enhancement, image segmentation, and feature extraction. Table 2.5 reports on image processing methods of the reviewed studies. Among all the thirty-three papers reviewed, thirty-two (97%) of them employed image processing methods for tasks such as feature extraction [55] and image segmentation [30] of archaeological ceramics. The remaining paper, ID33 [3], did not provide specific information on any image processing method.

2.4.1 Overview of Image Processing Methods

Image processing is crucial in the classification of ancient ceramic images and in the reviewed studies, including three main image processing approaches: feature extraction, image segmentation, and image enhancement.

Regarding feature extraction, an image feature is a primitive attribute of an image. Some features are natural and they are defined by the visual appearance of an image, while others are artificial features that result from specific manipulations of an image. Natural features include the luminance of a region of pixels and gray-scale textural regions [84]. Image amplitude histograms and spatial frequency spectra are examples of artificial features. Image features are important in the isolation of regions of common property within an image (image segmentation) and the subsequent identification or labelling of such regions (image classification).

Segmentation of an image involves dividing or separating the image into regions with similar attributes. The most basic attribute for segmentation is image luminance amplitude for a monochrome image and colour components for a colour image. Image edges and texture are also useful attributes for segmentation [31].

Image enhancement is a collection of techniques designed to improve the visual appearance of an image. For example, an image enhancement system might use high-

Table 2.5. Image processing of the reviewed studies

ID	Tasks	Image Processing Methods
ID1 [55]	Feature Extraction Image Segmentation	Convolutional neural network (CNN) used for feature extraction of porcelain patterns while Otsu's method is employed for image segmentation.
ID2 [69]	Feature Extraction	This study used the CNN for feature extraction of potteries.
ID3 [100]	Feature Extraction Image Segmentation	This study used the gradient vector flow field (GVF), soft contour information, and LBP operator to measure spatial coherence, contour continuity, and texture similarity of porcelain ewer images. Principal Component Analysis (PCA) was used to reduce the dimensionality. The segmentation was based on the SLIC-Neut.
ID4 [118]	Feature Extraction	This study used the CNN for feature extraction of porcelain vases.
ID5 [26]	Feature Extraction Image Segmentation	This study used gray level co-occurrence matrix (GLCM) to extract texture features of archaeological ceramic surfaces. Employed the CNN for feature extraction of ceramics. Used the watershed segmentation algorithm to segment the ceramic decoration images.
ID6 [30]	Image Segmentation	Four CNN models pyramid scene parsing network (PSPNet), DeepLabv3+, dual attention network (DANet), high-resolution network (HRNet) to segment porcelain patterns: dragon, fish, lotus, and peony.
ID7 [76]	Feature Extraction	GLCM and average Euclidean distance of Tamura texture features used as texture features of porcelains. Colour histogram and hue, saturation, value (HSV) colour space average similarity rate as colour features.
ID8 [56]	Image Preprocessing Image Enhancement Feature Extraction	A color segmentation algorithm is employed as part of the image preprocessing strategy, converting the color space from red, green, and blue (RGB) to HSV. FCutMix was used as an image enhancement method based on the CutMix method. Features extracted by two CNN models (ResNet) of different depths at different stages.
ID9 [10]	Feature Extraction	Variational autoencoder (VAE) used to compress images into a lower-dimensional space and then reconstruct them, thereby learning to extract the essential features from the pottery profiles.
ID10 [48]	Feature Extraction	Prewitt directional derivative approximation operator used to detect the edges of the copied image, Zernike moment discretization for feature extraction.
ID11 [19]	Image Preprocessing Feature Extraction	Data augmentation was achieved by a series of image preprocessing methods, rotation, translation, adjustment of chromatic aberration, amplification, distortion, and increase of noise. CNN used for feature extraction.
ID12 [47]	Image Enhancement Feature Extraction	Image enhancement included rotation, width shift, and zoom. Employed five CNN models (DenseNet121, InceptionV3, VGG16, GoogLeNet, and AlexNet) for feature extraction of patterns from ceramics.
ID13 [116]	Not provided	Not provided
ID14 [79]	Feature Extraction	Employed two CNN models(VGG16, ResNet50) for feature extraction of patterns from potteries
ID15 [117]	Image Segmentation Feature Extraction	Mask R-CNN used for segmentation to classify the ceramic form and generate the mask of the ceramic colours. Applied CNN model Inception V2 to extract features of ceramics according to their forms, materials, and patterns.
ID16 [61]	Feature Extraction	This study used VGG16 for feature extraction of porcelain bottles
ID17 [33]	Image Preprocessing Image Segmentation	Image quality improved using geometric distortion correction, image enhancement, and smoothing filtering. Edge detection is employed to define the shape and contours of the ceramics; and threshold segmentation for differences in grey characteristics between the ceramics and the background to produce a binary image.
ID18 [60]	Image Enhancement Feature Extraction	Image enhancement including random rotation, horizontal and vertical flipping, and random image brightness adjustments. Employed two CNN models (VGG19, ResNet50) for feature extraction of pottery fabrics.
ID19 [17]	Image Segmentation Feature Extraction	Image Segmentation for segment pore regions. Applied Hough Transform for feature extraction to investigate pores of potteries. Employed CNN model (GoogleNet) for feature extraction of segmented images
ID20 [119]	Image Preprocessing Feature Extraction	Image preprocessing including grayscale processing, ROI region detection, size normalization, filter denoising for improve quality of patterns. Applied LBP and HOG fusion feature extraction of banana leaf patterns.
ID21 [73]	Feature Extraction Image Segmentation	Eight-chain code method combined with a watershed algorithm is used for the shape extraction. The texture feature was extracted by Tamura method. Inscriptions segmented into single handwritten Chinese characters.
ID22 [126]	Image Segmentation Feature Extraction Image Registration	Extract curve patterns from a sherd using an FCN-based image segmentation method. A dual-source CNN was applied for feature extraction. Chamfer matching algorithm to match the pottery sherds to find their underlying designs
ID23 [34]	Feature Extraction	Combination of three texture features extraction methods: i) Gabor filters for local and oriented ones; ii) Laws' texture for levels, edges, spots, waves, and ripples; iii) Haralick's texture features for the distribution of grey levels.
ID24 [8]	Image Registration Feature Extraction	Image registration includes template matching algorithms to search for a specific pattern within a larger set of data according to the variations in scale, rotation, and colour. Employed deep neural network for feature extraction.
ID25 [44]	Feature Extraction	Applied weighted neighbour distance using Compound hierarchy of algorithms representing morphology to extract features from pottery images including statistical, textural, and pattern features.
ID26 [63]	Color Space Transformation Feature Extraction	Ceramic fragments images converted from RGB to HSV colour. The images were then divided into six equal-sized sub-blocks and the features extracted from these sub-blocks were based on the HSV values.
ID27 [127]	Feature Extraction	Applied kernel mean shift clustering to extract features (colour, texture, and shape) of porcelain images.
ID28 [110]	Feature Extraction	Employed LBP to extract texture features; A multi-channel colour histogram in the HSI (Hue, Saturation, Intensity) colour space is utilized to characterize the glaze colours of the porcelains.
ID29 [85]	Color Space Transformation Feature Extraction	The images of ceramic fragments were converted from RGB to HSV colour space. GLCM was applied to extract the texture feature of ceramic images.
ID30 [105]	Feature Extraction	Gabor filters were applied for extraction of descriptors while Bag of Visual Words (BoVW) extracted visual words based on the local image descriptors.
ID31 [80]	Feature Extraction	BoVW extracted features on the local image descriptors; colour and texture features were extracted from pottery images.
ID32 [3]	Feature Extraction	Pyramid histogram of visual words (PHOW), which is an extension of BoVW mode by extracting local features from different levels of a pyramid structure and then quantizing these features into a visual vocabulary.
ID33 [65]	Feature Extraction Image Segmentation	LBP was used to capture the texture information by analyzing the local spatial pattern of the pixels. Kirsch edge detection was used to detect edges in the sherds.

frequency filtering to emphasize the edges of objects in an image [84]. In this application, the image enhancement processor would emphasize salient features of the original image and simplify the processing task of data extraction.

2.4.2 Focus on Feature Extraction, Image Segmentation, and Image Enhancement

Depending on the objectives of the papers, different image processing methods are used. Primarily, they regard feature extraction, image segmentation, image pre-processing, image enhancement, colour space transformation, and image registration. Figure 2.7 shows the frequency of image processing tasks in the reviewed literature. In some studies, there is not a single task, but a combination of them.

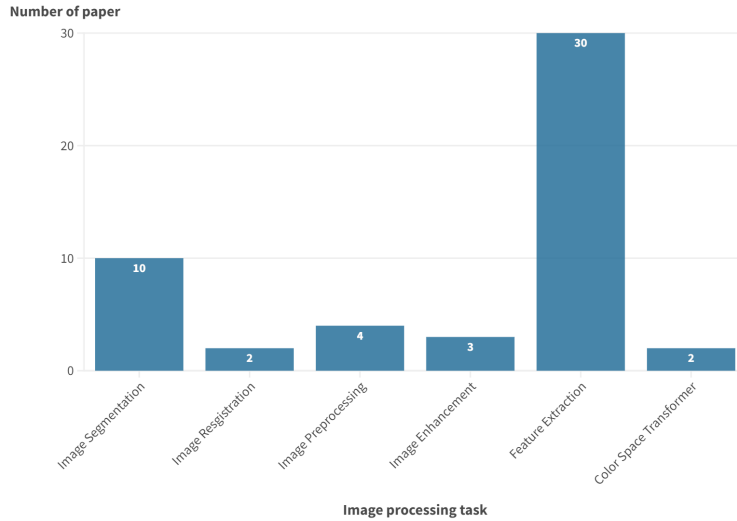


Figure 2.7. Frequency of Image Processing Tasks in Reviewed Literature

Thirty papers (93.7%) employed a feature extraction method, highlighting the central role of feature extraction in machine learning applications [55]. Different studies took advantage of several feature extraction methods according to different ceramic types (pottery and porcelain). Figure 2.8 shows the distribution of feature extraction methods in reviewed papers. These methods include traditional approaches, such as GVF and LBP to measure the counter of porcelains (ID3 [100]), GLCM to extract texture features, and HSV to extract color features (ID7 [76]), and Tamura texture feature (ID21 [73]). In addition to traditional feature extraction methods, there are shallow ML-based methods. ID27 [127] applied Kernel mean shift clustering to extract features (color, texture, and shape) of porcelain images. Both ID30 [105] and ID31 [80] used BoVW to extract

features from pottery images. Finally, thirteen papers employed deep neural networks for feature extraction including VAE model[10] and CNN model such as ResNet [56], AlexNet [47], VGG16 [61]), etc. From Figure 2.8, it can be seen that the CNN model for feature extraction is the most used method since DL has proven to be effective in feature extraction [55] and CNN represents the state-of-the-art in many image-related tasks [49].

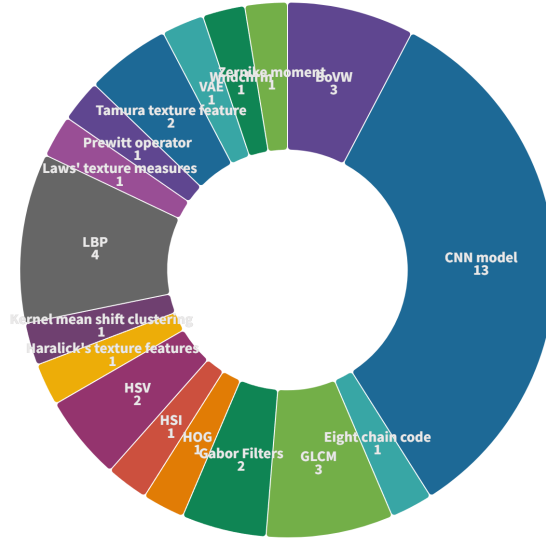


Figure 2.8. Distribution of Feature Extraction Methods in Reviewed Papers

Ten papers employed image segmentation techniques. Figure 2.9 shows the distribution of such methods. ID1 [55] used Otsu's method for image segmentation of porcelains, ID3 [100] applied SLIC-Neut algorithm to segment porcelain ewers. Finally, ID6 [30] used four CNN models to segment porcelain patterns while ID15 [117] employed Mask R-CNN for ceramics segmentation.

Four studies discussed image preprocessing and employed different methods for this task. ID8 [56] employed a color segmentation algorithm for image preprocessing strategy, ID11[19] used translation, adjustment of image chromatic aberration, ID17 [33] applied geometric distortion correction and smoothing filtering, and ID20 [119] used grayscale processing for improve quality of patterns.

Finally, three papers employed image enhancement techniques. ID8 [56] used FCutMix which is an image enhancement method based on DL, ID12 [47] and ID18 [60] used traditional enhancement methods including random rotation, horizontal and vertical flipping. Two papers indicated image registration such as Chamfer matching

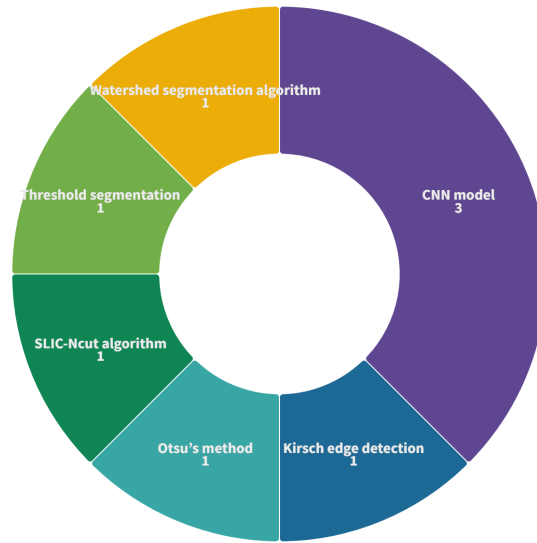


Figure 2.9. Distribution of Image Segmentation Methods in Reviewed Papers

algorithm [126] and template matching algorithm [8] for matching the pottery sherds or patterns. The other two papers (ID26 [63] and ID29 [85]) discussed color space transformation to convert RGB color format to HSV color format.

2.4.3 Evolving Landscape of Image Processing Methods

The previous Subsection highlighted that CNNs are the most used algorithms for both feature extraction and image segmentation. Figure 2.10 shows the evolving trend in the use of CNNs for these two tasks. Especially from 2021, CNNs became the dominant method for image processing in the field of archaeological ceramic images.

Such a shift in terms of popularity from traditional image processing methods (LBP, GLCM, Tamura texture feature, etc.) and traditional ML-based methods (Kernal mean shift clustering) to DL ones, reflects the same behaviour that can be observed in other research areas based on image analysis. In fact, traditional approaches rely on well-defined features of ceramics, and manual feature design is considered time-consuming and inflexible. On the other hand, DL overcome this limitation of handcrafted feature processing [55].

Although CNN's feature extraction method improves the accuracy of extraction, feature extraction of ancient ceramics still faces some challenges, including the inconsistency of surface quality and the complexity and diversity of texture [26]. A promising research direction that could be exploited regards the use of Transformers for feature extraction.

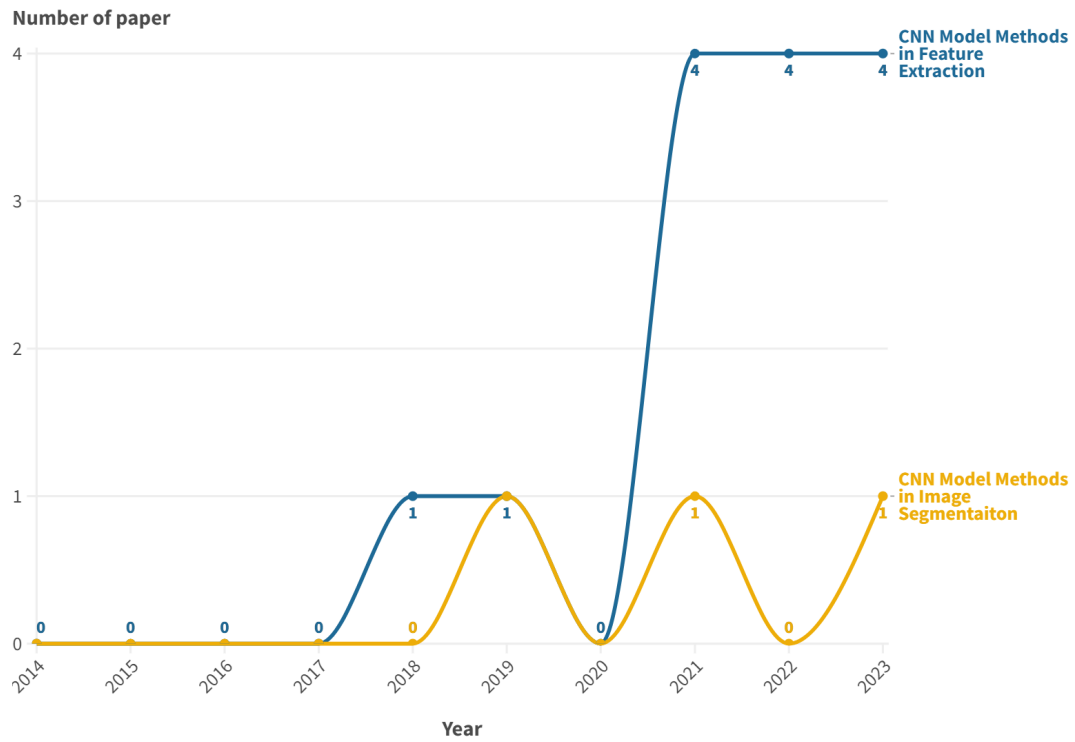


Figure 2.10. Trend of CNN Model Methods in Feature Extraction and Image Segmentation

In fact, parallel to what is happening in computer vision [122], Transformers could also be exploited in image analysis of archaeological ceramics. These models excel at capturing long-range dependencies within data [106], a crucial aspect for analyzing complex visual details in ceramics. Unlike traditional methods that focus on local features, transformers can identify subtle relationships between different parts of the image, potentially revealing characteristics like decorative patterns, specific glazing techniques, or even signs of wear and tear. This deeper understanding of the image could lead to a more accurate identification of ceramic origins, types, and potential historical context. Furthermore, transformers can be trained on diverse datasets encompassing a wide range of ceramics, potentially overcoming limitations caused by data scarcity in this field.

2.5 Machine Learning Algorithms for Porcelain Classification (RQ1.3)

This section answers the third and final research question **RQ1.3**, "What are the main ML algorithms used in the identification of archaeological ceramics?" The algorithms employed for the identification of archaeological ceramics mainly consist of classification and object detection algorithms [64]. Table 2.6 reports on the various ML algorithms employed. Among thirty-three papers, twenty-eight (84.8 %) of them use ML algorithms. Among these twenty-eight papers employing ML algorithms, seventeen of them (60.7%) used DL algorithms to identify archaeological ceramics, using various CNN architectures: VGG16, ResNet50, DenseNet121, InceptionV3, GoogLeNet, and AlexNet [19, 17, 79, 61]. The papers employing traditional ML algorithms to classify archaeological ceramics took advantage of SVM, KNN, and SOM [44, 63, 85].

The remaining ones (15.1%), instead, are primarily focused on datasets and image processing methods of the archaeological ceramics and did not conduct experiments for the classification or identification of archaeological ceramics.

2.5.1 Application of ML Algorithms in Archaeological Ceramic Images

Figure 2.11 shows the distribution of applications based on ML/DL algorithms in reviewed papers. Among twenty-eight papers, twenty-five of them applied ML/DL algorithms to image classification, three of them focused solely on image segmentation, and two of them discussed image processing and object detection, respectively. It is worth noting that some papers have more than one application, combining image segmentation with another application. Anyway, these distributions highlight that the problem of archaeological ceramic identification is often modelled and tackled as a classification one.

Fourteen studies employed a DL algorithm, such as VAE [10] and CNN like VGG [69], ResNet [118], EfficientNet [69], MobileNet [118], GoogleNet [47], etc., while thirteen studies took advantage of an ML algorithm, including [85], SVM [44], SOM [63], KNN [85], etc.

Although there is a similar amount of reviewed literature based on ML and DL, they were used in different periods. Figure 2.12 shows the tendency of ML/DL algorithms for image classification. There's a notable increase in the use of DL algorithms starting around 2022, as indicated by the sharp upward line, while ML algorithms trend line shows fluctuations over the years, with peaks and troughs, indicating variable usage or effectiveness during that period. It is interesting to notice that after 2021, ML

Table 2.6. ML algorithms of the reviewed studies

ID	Application		ML Algorithms
ID1 [55]	Image	Classifica- tion	This study employed a convolutional neural network (CNN) as a primary tool for porcelain image classification with an accuracy of 94.2%.
ID2 [69]	Image	Classifica- tion	The study employed CNN for image classification of potteries. Evaluated the performance of three different CNN architectures: Visual geometry group (VGG), Residual neural network (ResNet), and EfficientNet, achieving accuracy of 66%, 77%, and 79%, respectively.
ID3 [100]	Image	Classifica- tion	This study employed logistic regression cross-validation (CV) for the image classification with an accuracy of 90.7%.
ID4 [118]	Image	Classifica- tion	The study employed CNN for image classification of porcelain vases. Compared and evaluated the accuracy and loss values under different optimization parameters under three different CNN models: ResNet, Inception, and MobileNet, achieving an accuracy of 93.8%, 91.2%, and 94.2%, respectively.
ID5 [26]	Image	Classifica- tion	The study employed CNN to extract texture features of ceramics for image classification with an accuracy of 76.3%.
ID6 [30]	Image	Segmenta- tion	This paper was based on four CNN models PSPNet, DeepLabv3+, DANet, HRNet to segment porcelain patterns: dragon pattern, fish pattern, lotus pattern and peony pattern. Results show that the image segmentation methods of HRNet and DANet models are better than the other two models.
ID7 [76]	Not provided		Not provided
ID8 [56]	Image	Classifica- tion	This study employed a Feature Fusion Convolutional Network (FFCNet), this architecture extends the typical capabilities of the standard CNN (ResNet) by incorporating mechanisms specifically designed for effective feature fusion, improving the recognition and classification accuracy from 71.7 to 79.5%.
ID9 [10]	Image Processing		Convolutional VAE was designed to work in an unsupervised learning style, suitable for this study where labeled data was not available. VAE extracted the essential features of archaeological pottery profiles. The model demonstrated its effectiveness and was validated through visual inspection of the reconstructed images against the original profiles.
ID10 [48]	Not provided		Not provided
ID11 [19]	Image	Classifica- tion	This paper employed CNN for image classification of porcelain vases, accuracy of the network in identifying the types and ages of ancient ceramics was above 96.4%.
ID12 [47]	Image	Classifica- tion	This study employed and compared the performance of five CNN models(DenseNet121, InceptionV3, VGG16, GoogLeNet, and AlexNet), the comparison of the CNN performance showed that the VGG16 achieved the best results with accuracy of 86.5%.
ID13 [116]	Image	Classifica- tion	This study compared ML algorithm logistic regression for DL algorithm based on transfer learning, and found that the accuracy of DL algorithm was 70% and the accuracy of logistic regression was 48%, the DL algorithm improved classification results by more than 20% over the logistic regression.
ID14 [79]	Image	Classifica- tion	This research presented the evaluation for image classification of decorated potteries using two CNN models(VGG16 and ResNet50) with an accuracy of 82.5%.
ID15 [117]	Image	Segmenta- tion Object Detection	This article employed Mask R-CNN for segmentation to classify the ceramic form and generate the mask of the ceramic, the average accuracy of Mask R-CNN was 79.5%. For the analysis of material and pattern, the Faster R-CNN was applied by using the region proposal networks (RPN), the average accuracy of Faster R-CNN was 60%.
ID16 [61]	Image	Classifica- tion	This paper compared two different models for image classification: a DL model (VGG16) with an accuracy of 83.9%, an integration of DL and ML model (VGG16-GSSVM) with an accuracy of 89.3%
ID17 [33]	Not provided		Not provided
ID18 [60]	Image	Classifica- tion	This research presented the evaluation for image classification of pottery fabrics in thin sections using two CNN models (VGG19 and ResNet50) and the accuracy remained above 93%.
ID19 [17]	Image	Classifica- tion	This study compared ML method SVM with DL method CNN model (GoogleNet) for image classification. The CNN method(higher than 90%) showed more accurate than the SVM one(less than 80%).
ID20 [119]	Image	Classifica- tion	This study employed SVM combined with LBP-HOG fusion feature extraction for the porcelain image classification of three dynasties, achieved 75% of Yuan dynasty, 80% of Ming dynasty and 90% of Qing dynasty.
ID21 [73]	Not provided		Not provided
ID22 [126]	Image	Segmenta- tion Image Classifica- tion	This paper used an FCN-based image segmentation method to extract curve patterns from a sherd A dual-source CNN (AlexNet) was applied for feature extraction, and the segmented curve patterns could be implicitly classified through this method.
ID23 [34]	Image	Classifica- tion	This paper employed k-Nearest Neighbours (KNN) algorithm for image classification, compared three different methods for texture extraction, Gabor filters with an accuracy of 68.1%, Laws' texture measures with an accuracy of 73.9%, Haralick's texture features with an accuracy of 71.1%.
ID24 [8]	Image	Classifica- tion	This study employed a deep neural network for image classification of ceramics.
ID25 [44]	Image	Classifica- tion	The classification in Wndchrn involves different ML algorithms(decision trees, SVM, or simple statistical comparisons), integrated within its framework to classify pottery images. The classification accuracy reached between 90% to 95%.
ID26 [63]	Image	Classifica- tion	This paper applied a Self-Organizing Map(SOM), a type of unsupervised ML model to group the fragments with an accuracy of 89.6%.
ID27 [127]	Image	Classifica- tion	This paper employed semi-supervised mean shift clustering, an ML method for image classification with an accuracy of 75.6%
ID28 [110]	Not provided		Not provided
ID29 [85]	Image	Classifica- tion	This paper employed the KNN algorithm for image classification with an accuracy of 86.5%.
ID30 [105]	Image	Classifica- tion	SVM was used to classify different patterns of pottery shards using features extracted from Gabor filters and features generated by Bags of visual words. Results of classification were presented by a confusion matrix.
ID31 [80]	Image	Classifica- tion	Compared five different ML algorithms for classification, SVM, KNN, Naive Bayes, Sequential Minimal Optimization (SMO) and Simple Logic, SVM achieved the accuracy of 93.96%.
ID32 [3]	Image	Classifica- tion	Employed two ML algorithms, KNN and SVM on image classification of patterns from ceramics.
ID33 [65]	Image	Classifica- tion	KNN and SVM used for the classification of patterns from pottery sherds.

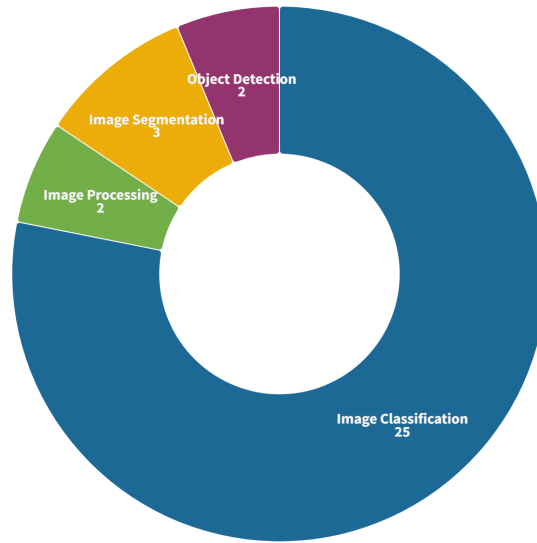


Figure 2.11. Distribution of Applications in Reviewed Papers

algorithms are mainly used as baseline models for comparison. Hence, we can conclude that DL algorithms for the classification of archaeological ceramic images have been the dominant trend in the last three years.

All three papers on applications of image segmentation are based on the DL algorithm, including ID6 [30] compared four CNN segmentation models on porcelain patterns, ID15 [117] used Mask R-CMM for segmentation, and ID22 [126] employed FCN model for segmentation of pottery sherds.

Finally, the only paper about object detection is also based on the DL algorithm. It combines object detection with image segmentation, proposing a new tool to allow users to analyze the visual elements of ceramics using their smartphones and devised a method to extract representative colors of ceramics for object detection.

2.5.2 Performance Evaluation

The performance evaluation of algorithms for image classification, image segmentation, and object detection is mainly based on evaluation metrics including accuracy, precision, recall, and F1 score [117].

With regard to image classification, the papers based on CNNs architectures perform differently on various datasets. In the reviewed literature, the performance evaluation of the models is mainly based on accuracy. Table 2.7 reports the accuracy of DL algorithms in the classification of ancient ceramics. It can be seen that the classification accuracy is more than 70% in both pottery and porcelain, and the accuracy of ID1, ID4, ID11, ID18,

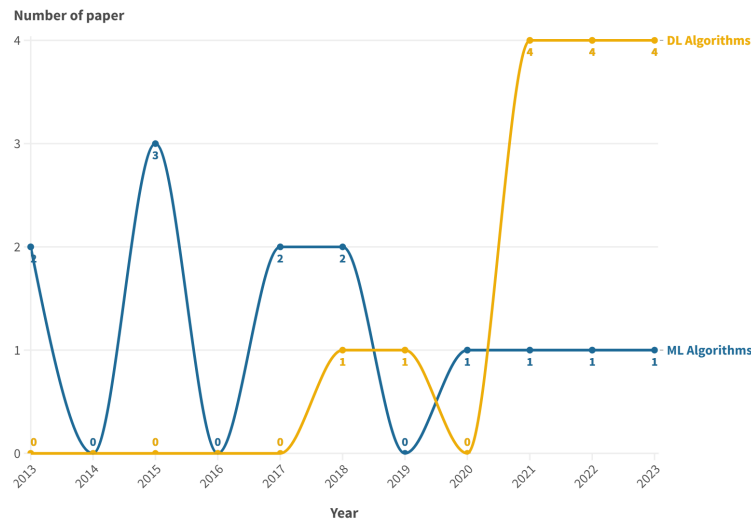


Figure 2.12. Tendency of ML/DL Algorithms on Image Classification

and ID19 even exceeds 90%. The achieved results are in line with the state-of-the-art image classification using deep learning algorithms.

Table 2.8 reports the accuracy of ML algorithms in image classification. Seven papers employed SVM and achieved accuracy higher than 70%. In ID13 and ID19, ML and DL algorithms were applied separately to image classification to compare the results. They showed that the accuracy of DL algorithms was higher than the one achieved by ML algorithms. Although SVM has achieved good classification results as an ML algorithm, for example on pottery shards, its performance was still inferior to that of the DL algorithm on the same dataset [17]. Therefore, the choice of the DL algorithm as an image classification of archaeological ceramics appears to be a natural choice.

2.6 Discussion and Conclusion

To address the traditional limitations of systematic reviews, we followed the PRISMA methodology in this work. However, this approach still presents several limitations. Firstly, the review process was highly selective, focusing only on scientific articles published in journals or conference proceedings from reputable databases to ensure high-quality content. Secondly, the study of citations or reference analysis was deliberately excluded in order to keep the chapter more concise and focused. However, future research efforts aimed at a more comprehensive overview could expand the range of review sources and employ knowledge graphs to visually describe developing tendencies and research

Table 2.7. Accuracy of DL algorithms on image classification

ID	Ceramic Type	Number of Classes	DL Algorithms	Accuracy
ID1 [55]	Porcelain (Yaozho kiln)	N/A	CNN model	94.2%
ID2 [69]	Pottery	7 classes	VGG ResNet EfficientNet	66% 77% 79%
ID4 [118]	Porcelain	19 types of vases	MobileNet ResNet Inception	94.2% 93.8% 91.2%
ID5 [26]	Porcelain	N/A	CNN model	76.3%
ID8 [56]	Porcelain	14 types	ResNet FFCNet	71.7% 79.5%
ID11 [19]	Porcelain	2 types of vases	CNN model	96.4%
ID12 [47]	Ceramic (Sukhothai kiln)	7 types of patterns	VGG16	86.5%
ID13 [116]	Porcelain	3 glaze colours	CNN model	70%
ID14 [79]	Pottery	8 design classes	VGG16, ResNet50	82.5%
ID16 [61]	Porcelain	7 types of bottles	VGG16 VGG16- GSSVM	83.9% 89.3%
ID18 [60]	Pottery	5 types	VGG19, ResNet50	93%
ID19 [17]	Pottery	2 types of patchwork	CNN model	90%
ID22 [126]	Pottery	98 non-composite patterns 20 composite patterns	AlexNet	N/A
ID24 [8]	Ceramic	7 classes	CNN model	N/A

Table 2.8. Accuracy of ML algorithms on image classification

ID	Ceramic Type	Number of Classes	ML Algorithms	Accuracy
ID3 [100]	Porcelain	3 dynasties	Logistic regression cross-validation	90.7%
ID13 [116]	Porcelain	3 glaze colours	Logistic regression	48%
ID19 [17]	Pottery	2 types of patchwork	SVM	80%
ID20 [119]	Porcelain	3 dynasties	SVM	75%
ID23 [34]	Pottery	2 classes	KNN	68.1%
ID25 [44]	Pottery	2 types of clays	SVM	90%
ID26 [63]	Ceramic	8 classes of fragments	SOM	89.6%
ID27 [127]	Porcelain	13 classes	Mean shift clustering	75.6%
ID29 [85]	Porcelain	6 classes of fragments	KNN	86.5%
ID30 [105]	Pottery	6 classes	SVM	N/A
ID31 [80]	Pottery	N/A	SVM	93.96%
ID32 [3]	Ceramic	6 classes of fragments	KNN, SVM	N/A
ID33 [65]	Pottery	2 classes of sherds	KNN, SVM	N/A

hot spots in the application of machine learning algorithms to archaeological ceramics identification in a more sophisticated way. Thirdly, this literature also had unaddressed research areas, first was integration with other technologies, the potential integration of DL algorithms with other technologies such as 3D scanning and Virtual Reality(VR) for more comprehensive analysis of archaeological ceramics, second was cross-cultural comparative studies, investigating how DL algorithms can be used to the identification of archaeological ceramics across different cultures and historical periods more effectively. Finally, since this research began at the beginning of 2024, it includes reviewed studies that have been published during the last ten years (i.e., between 2013 and 2023). The limitation of this timeframe indicates that future works could combine studies from 2024 and subsequent years, which will have the potential to provide deeper insights and more varied applications for the review of archaeological ceramics.

The purpose of this chapter is to examine the trends of machine learning in the identification of archaeological ceramics. Specifically, it focuses on the application of ML and DL algorithms between 2013 and 2023, on several types of ceramic (pottery and porcelain). It investigates several aspects such as the available datasets, the image processing methods, and ML or DL algorithms.

In terms of the application of ML algorithms to the image classification of ceramic images, papers from 2013 to 2020 tend to use traditional ML algorithms such as SVM and KNN. However, there is a general shift towards the use of DL based on CNN models for classification in papers from 2020 to 2023. In fact, based on the conclusions drawn from these papers, DL algorithms have made significant breakthroughs in many tasks due to their powerful automatic feature extraction capabilities. Among CNN models, VGG16, MobileNet and ResNet have achieved good performance on archaeological ceramic images and higher accuracy compared to traditional ML algorithms such as SVM and KNN. This indicates that DL algorithms such as CNN models will be the main trend in the future for archaeological ceramic classification applications.

Various studies demonstrated the technical potential of DL for the identification of archaeological ceramics. The manual classification of a large number of pottery sherds is a time-consuming activity and empirical identification methods of porcelains often lack objective constraints. The image classification based on DL provides support for the classification of archaeological ceramics. This review and its conclusions should be seen as a first step in the analysis of machine learning for ceramic identification in the sector.

Overall, the use of machine learning algorithms in archaeological ceramics classification complements traditional methods by offering efficiency, accuracy, and new

avenues for data analysis and interpretation. By combining the strengths of both approaches, researchers can gain deeper insights into past civilizations and cultural practices. The efficiency of these algorithms can speed up the classification process allowing archaeologists to analyze larger datasets and automate repetitive tasks, such as sorting and categorizing artifacts based on specific attributes. This automation frees up archaeologists' time to focus on higher-level analysis and interpretation. Moreover, machine learning algorithms can uncover hidden relationships or correlations within ceramic datasets that may not be apparent through traditional methods. By analyzing these insights, archaeologists can gain a deeper understanding of cultural practices, trade networks, and historical contexts.

Chapter 3

Evaluating CNN Effectiveness in Multi-Task & Multi-Label Porcelain Classification

To what extent can deep learning models (especially CNN-based multi-task architectures) accurately classify key attributes of Chinese porcelain under real data constraints?

What performance can be achieved when simultaneously classifying dynasty, kiln provenance, glaze, and vessel form?

How does transfer learning influence classification performance under limited, imbalanced archaeological datasets?

How can explainable AI methods help interpret model decisions and relate computational features to archaeologically meaningful attributes?

— RQ-2, RQ-2.1, RQ-2.2, RQ-2.3

This chapter explores the performance of these CNN architectures in classifying porcelain artifacts across multiple tasks: dynasty, glaze, kiln, and type. It presents a comparative analysis of models, evaluating their performance. Furthermore, the impact of transfer learning is evaluated by comparing models trained with and without pre-trained weights. Additionally, we incorporated eXplainable Artificial Intelligence (XAI) techniques to provide valuable insights into the models' decision-making processes.

3.1 Introduction and Background

More recent studies have demonstrated the effectiveness of DL, particularly CNNs, in the feature extraction and classification of porcelain images. Yang et al. [116] employed a custom CNN for ceramic type recognition, achieving an accuracy of 70%. They compared the performance of their DL approach against a traditional ML approach (i.e., logistic regression). Also Ma et al. [61] focused on type recognition of ancient ceramics. They proposed a novel identification method called VGG16-GSSVM, which is specifically designed for few-shot learning. This approach uses a pre-trained VGG16 model to extract shape features and then employs an SVM classifier, with its parameters optimized using a grid search algorithm. The experiment demonstrated that the VGG16-GSSVM model surpassed the accuracy and efficiency of traditional CNNs and CNN-SVM models, reaching an accuracy of 89%. Always in the context of porcelain type classification, Yue et al. [118] compared the performance of Inception, ResNet, and MobileNet, three well-known CNNs. The study found that all three models were highly effective at recognizing and classifying ceramic types, achieving an accuracy of up to 90%.

Beyond type recognition, researchers have also leveraged DL to tackle a variety of other problems in ceramic analysis. Liu [55] focused on ware recognition, specifically focusing on Yaozhou kiln ceramics. The author employed a custom CNN identifying and classifying ceramics. In the end, the CNN achieved 94.2% accuracy. Instead, Lyons in [60] addressed the classification of ceramic fabrics. He compared two CNNs, VGG19 and ResNet50, using a transfer learning methodology. The CNNs were employed to automatically recognize and classify ceramic fabrics from images of petrographic thin sections. By leveraging models pre-trained on the ImageNet dataset, the authors were able to achieve over 93% accuracy in classifying thin section images into five fabric groups. A different task was addressed by Gao et al. [26], where he focused on ceramic texture extraction. He took advantage of a custom CNN, integrating image processing techniques like edge detection and thresholding. He achieved an average AUC score of 0.93. Finally, Mendonca et al. [69] tackled the problem of ceramics material recognition by analyzing potsherds' X-ray images. The authors compared the performance of three CNNs VGG11, ResNet-152, and EfficientNetB7, which achieved a 90% classification rate. Unlike ML, DL models automatically learn hierarchical features directly from raw images, capturing complex visual characteristics like texture, glaze patterns, and shapes. However, DL methods demand substantial computational resources and large datasets, which can be a challenge in the context of cultural heritage research.

Multi-task learning leverages shared representations across multiple related tasks,

leading to improved data efficiency, faster learning, and better generalization [28]. This technique is particularly beneficial for archaeological porcelain classification, where multiple attributes, such as dynasty, ware, glaze, and type, can be learned concurrently from the same dataset. By sharing feature representations during training, the model optimizes performance across all classification tasks simultaneously. Transfer learning has emerged as a powerful approach to address the challenges posed by limited datasets. Studies by Liu et al. [56] and Yang et al. [116] demonstrated the effectiveness of transfer learning using CNN models pre-trained on large datasets like ImageNet. This strategy exploits the general feature representations learned from extensive image datasets, significantly reducing the need for large labelled datasets in specialized domains like porcelain classification. Tajbakhsh et al. [103] further highlighted that the early layers of CNNs learn generic patterns, such as edges and textures, which are transferable across diverse image domains. Two primary transfer learning strategies are prevalent: feature extraction, where pre-trained models serve as fixed feature extractors, and fine-tuning, where pre-trained weights are partially or fully retrained on the target dataset. This chapter explores both approaches using five state-of-the-art architectures: ResNet50, ResNet101, MobileNetV3, EfficientNetB2, and InceptionV3 [51]. Additionally, we incorporated eXplainable Artificial Intelligence (XAI) techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM) and a SHapley Additive exPlanations (SHAP) inspired method, to move beyond simple performance metrics and provide valuable insights into the models' decision-making processes by visualizing and quantifying the features they rely on for classification.

This chapter explores the performance of different CNN architectures in classifying porcelain artifacts across multiple tasks: dynasty, glaze, kiln, and type. structured as follows: Section 3.2 describes the porcelain dataset collection process and label configuration standard according to the traditional morphological authentication of porcelain. Section 3.3 outlines the selection of CNN architectures and the design of multi-task learning frameworks. Section 3.4 presents a comprehensive evaluation of all model architectures on the test set, analysing both task-specific and overall multi-task performance. Section 3.5 discusses this chapter by concluding.

3.2 Dataset and Label Configuration

3.2.1 Data collection and collation stage

Due to the diversity and complex system of archaeological Chinese porcelains, data collection requires multiple sources to cover dynasties, kilns, glazes and object types. In

this part of the thesis, we focus on the porcelain dynasties from the Song (960–1279 CE) and Yuan (1271–1368 CE) dynasties. Sample images of porcelains are collected from the open data of the National Palace Museum in Taipei and the Palace Museum in Beijing, as described in Table 3.1.

Table 3.1. Image Sources and Licences for Museum Digital Collections

Code	Museum	Licence
PMBJ	Palace Museum, Beijing	Images obtained through https://www.dpm.org.cn/explore/collections
PMTP	National Palace Museum in Taipei	CC-BY-4.0; images obtained through https://digitalarchive.npm.gov.tw/opendata

After eliminating the images that did not meet the experimental requirements, we retained 7,263 high-resolution porcelain images from the Song and Yuan dynasties, reported in Table 3.2. Each image was captured under controlled lighting conditions with standardised colour calibration to ensure consistency across different sources. This dataset includes different glazes, colours, kilns and object types. Period attributions and related metadata were provided by archaeologists affiliated with the contributing museums and are publicly available on their websites, ensuring both authenticity and diversity of sources. Finally, culturally sensitive artefacts were masked before publication, and all licences were verified to permit non-commercial academic use. No human-subject data were involved.

Table 3.2. Distribution of Song and Yuan dynasty porcelain images by data source

Source	Song	Yuan	Total
PMBJ	358	91	449
PMTP	5 304	1 510	6 814
Total	5 662	1 601	7 263

3.2.2 Historical and archaeological context

In order to understand the visual features that our CNNs must learn to recognise, we take a dynasty-based approach to provide a comprehensive description of the characteristics of Chinese porcelain. We examine each historical period through the following five key dimensions: period style, kilns, glazes, decorative patterns and vessel types. Figure 3.1 below provides the provenance of the porcelain kilns used in the dataset [20]. The

map illustrates how the excellence of Song (960-1279 CE) to Yuan (1271–1368 CE) porcelains was rooted in two major regions: the north-central region (Henan–Hebei) and the south-eastern region (Jiangxi–Zhejiang–Fujian).



Figure 3.1. Map with the provenance of the porcelain Kilns Used in the Dataset

Evolution of Chinese porcelain kilns: from Song diversity to Yuan centralization

The Song dynasty was a prosperous period for the development of Chinese porcelains, reflecting the social, economic, and cultural prosperity of the time [123]. Song porcelain is characterised by the presence of two major systems, as shown in Table 3.3. The first

of these is the official kiln system, which is frequently referred to as the Five Great Kilns (Ding, Ru, Jun, Guan and Ge) of the Song Dynasty. The second system is the folk kiln system, which is divided into eight kiln systems, exactly distributed along the Yangtze River, with four located in the northern region and four in the southern portion. In the northern region, the four kiln systems were located at the Cizhou Kiln, the Yaozhou Kiln, the Jun Kiln, and the Ding Kiln. In the southern region, the kiln systems were situated at the Jingdezhen Kiln, along with the Longquan Kiln, the Jian Kiln, and the Jizhou Kiln [62].

Table 3.3. Major Song-Dynasty Kiln Systems and Representative Sites

System	Kiln Name	Region*
Official kiln system (Five Great Kilns)	Ding	North
	Ru	North
	Jun	North
	Guan	South
	Ge	South
Folk kiln system (Eight regional kilns)	Cizhou	North
	Yaozhou	North
	Jun (folk) [†]	North
	Ding (folk) [†]	North
	Jingdezhen	South
	Longquan	South
	Jian	South
	Jizhou	South

* “North”/“South” are defined relative to the Yangtze River.

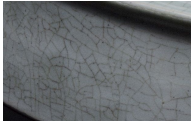
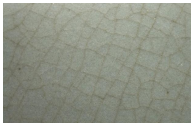
[†] Folk-kiln output that shares the name of an official kiln.

The following discussion will examine the official kiln system of the Song Dynasty, specifically the five great kilns that were constructed during this era.

Ru kilns are exceedingly uncommon and challenging to locate. The firing time is approximately 20 years, and due to the rarity, Ru kilns are of significant value. One of the characteristics of Ru kilns is the "crab claw pattern". The "crab claw pattern" is the presence of brown holes in the glaze, resulting from the fracture of glaze bubbles during the process of formation (Table 3.4). The second characteristic is the "fish scale crackles", and the angle of crackles exhibits an oblique orientation, as opposed to a straight orientation (Table 3.4). The third is the grey body, which exhibits a light grey, and the body is thin (Table 3.4). The fourth characteristic is that the glaze colour of the Ru kiln is bluish-green. The Ru kiln is plain and unadorned in general. Owing to

the thickness of the glaze, the body has been inadequately carved [62].

Table 3.4. Characteristics of Song Ru Kiln with Example Photos

Characteristic	Image	Description
Crab claw pattern		Detail of "crab claw pattern" on basin in celadon glaze of Ru kiln, presence of brown holes in the glaze.
Fish scale crackles		Detail of "fish scale crackles" on bowl in celadon glaze of Ru kiln, demonstrating the crackling pattern.
Grey body		Detail of Dish in bluish-green glaze of Ru kiln, showing the thin grey body of the bowl rim.

Note. All images are drawn from the dataset compiled for this study.

The Ge kiln is characterised by "gold wire and iron thread". The "gold wire" is characterised by its fine crackles, which are yellow; the "iron thread" is distinguished by its large crackles, which are black (Table 3.5). The "gold wire and iron thread" is a unique aesthetic quality attributed to the effects of the porcelain firing process. During the firing process, the different expansion coefficients of the body and glaze result in the glaze cracking. The second feature of the Ge kiln is the "purple rim and iron foot", which refers to the thin glaze that covers the rim, thereby revealing the purple body, and the iron-brown colour of the exposed foot (Table 3.5). Both of which are also found in the Guan kiln [62].

Table 3.5. Characteristics of Song Ge Kiln with Example Photos

Characteristic	Image	Description
Gold wire and iron thread		Detail of "gold wire and iron thread" on bowl in celadon glaze of Ge kiln, presence of fine and large crackles.
Purple rim and iron foot		Detail of "purple rim and iron foot" on plate in celadon glaze of Ge kiln, showing purple body of the bowl rim.

Note. All images are drawn from the dataset compiled for this study.

The colour of the Guan kilns is slightly more translucent than that of the Ru kilns,

appearing somewhat more greenish. The characteristic features of the Northern Song and Southern Song Guan kilns are their "purple rim and iron foot". Secondly, the Guan kilns have crackles. Thirdly, the kilns of the Northern Song Dynasty are characterised by their thickness and weight, in contrast to the slender and lightweight design during the Southern Song Dynasty. Despite the similarities between the Ge and Guan two kilns. Indeed, it remains feasible to discern the distinction. It is commonly observed that one should allocate a large piece of crackle for the Guan kiln and a small piece for the Ge kiln. Secondly, the Ge and Guan two kilns are part of the celadon system, and the glaze and crackles are regarded as forms of artistic expression. The objective is to embody both the aesthetic appeal of glaze colour and the visual allure of crackles. Thirdly, the Guan kilns and the Ge kilns represent the highest official aesthetic standard of the Song Dynasty [62].

A distinguishing characteristic of the Jun kilns is the use of a milky glaze, which is characterised by its opacity. It is evident that Guan kilns, Ge kilns and Ru kilns are all transparent vitreous glaze, and that the glaze is translucent. The colour and lustre of Song dynasty Jun porcelain is characterised by its diffusion, with the edges of the colours being indiscernible. The presence of trace amounts of copper oxide in the glaze has resulted in the successful creation of a distinctive purple-red glaze of Jun kilns. A flowerpot, for instance, is characterised by a uniform colour, typically a shade of rose purple. Following the Jin and Yuan Dynasties, the presence of discernible colour spots on porcelain became increasingly evident, as described in Table 3.6. These spots could be arranged into discernible shapes, thereby signifying an evolution from the Song to the Yuan era [62].

Table 3.6. Comparative Analysis of Jun Kiln Flowerpots: Song Dynasty versus Yuan Dynasty

Porcelain Title	Song dynasty Jun kiln flowerpot	Yuan dynasty Jun Kiln flowerpot
Image		
Glaze Colour	Rose purple, uniform color	Rose purple, visible color spots
Color Edges	Indiscernible edges	Discernible spot boundaries

Note. All images are drawn from the dataset compiled for this study.

Of the Five Great Kilns of the Song Dynasty, the Ding kiln is distinguished from the other four by the type of porcelain it produces, which is white porcelain. During the Song dynasty, the process of firing white porcelain was more challenging than the process of firing celadon. Except for a small number of exceptions, the other four kilns have not been carved. The underlying reason for this phenomenon is attributed to the disparity in the thickness of the glazes employed in the kiln's production. Specifically, the glaze of the Ding kiln is characterised by its remarkable thinness, in contrast to the thicker glazes utilised in the other four kilns. This distinction enables the Ding kiln to be carved with a knife. The white glaze has been decorated with three different types of pattern: printed, incised and carved, reported in Table 3.7. The printed decoration is executed through the use of moulds, while incised and carved decorations are performed with the aid of bamboo or metal tools. Wider lines with distinct knife edges characterise incised patterns, while carved patterns exhibit finer lines. The most prevalent motifs of these three types are floral patterns, with lotus patterns being the most common on incised vessels. Carved patterns feature floral motifs, animal motifs, and ripple motifs. A wide variety of floral motifs can be found in printed patterns, with flowers such as the lotus, chrysanthemum, peony, and over ten different varieties. Following the midpoint of the Song Dynasty, Ding kilns adopted a novel process known as "upside-down firing". The term "upside-down firing" is used to describe the process of buckling over to burn. The rim is to be formed into a coarse circle, without glazing. Consequently, the absence of glazing of the rim is a distinctive feature of the Ding kiln [62].

The folk aesthetics of the Song Dynasty differed from the official ones on account of the populace's greater pragmatism. Consequently, the folk art was heavily decorated. In the northern region, the four folk kiln systems were located at the Cizhou Kiln, the Yaozhou Kiln, the Jun Kiln, and the Ding Kiln. In the southern region, the folk kiln systems were situated at the Jingdezhen Kiln, along with the Longquan Kiln, the Jian Kiln, and the Jizhou Kiln [62].

The significant kiln in the northern region is the Cizhou kiln, which is decorated by three means. The first of these is the utilisation of hard tools, such as knives and bamboo, for carving. The second method involves the application of paint with a soft brush directly onto the white body. The third method involves the utilisation of both hard and soft brushes in conjunction with each other, as shown in Table 3.8. The raw materials utilised in the fabrication of the Cizhou kiln are inferior to the kaolin clay of Jingdezhen. The Cizhou kiln invented a craft known as "Slip". Following the completion of the initial porcelain, a layer of white powder is applied to the porcelain body, and subsequently, the glaze is applied. This layer of white powder is referred to as "Slip".

Table 3.7. Comparative Analysis of Song Ding White Glaze Decoration Techniques and Motifs

Decoration Technique	Printed	Incised	Carved
Image			
Porcelain Title	Bowl with printed chrysanthemum decoration in White glaze, Ding kiln (Song Dynasty)	Plate with incised lotus decoration in White glaze, Ding kiln (Song Dynasty)	Bowl with carved lotus decoration in White glaze, Ding kiln (Song Dynasty)
Tools Used	Moulds	Bamboo or metal tools	Bamboo or metal tools
Line Characteristics	Moulded formation	Wider lines with distinct knife edges	Finer lines
Primary Motifs	Various floral motifs: lotus, chrysanthemum, peony, over ten varieties	Floral patterns, lotus patterns most common	Floral, animal, and ripple motifs

Note. All images are drawn from the dataset compiled for this study.

The most common variety of Cizhou kiln is characterised by underglaze slip painting and is renowned for its strength. The features of the Cizhou kiln are principally frankness and roughness, which reflect the mood of the folk [62].

Table 3.8. Comparative Analysis of Song Cizhou Kiln Decoration Techniques and Characteristics

Decoration Technique	Hard Tool Carving	Soft Brush Painting	Combined Hard and Soft Tools
Image			
Porcelain Title	Ruyi-shaped Pillow with carved decoration in white glaze, Cizhou kiln (Song Dynasty)	Pillow with Underglaze black decoration over a white ground, Cizhou kiln (Song Dynasty)	Octagonal Pillow with Underglaze black decoration over a white ground, Cizhou kiln (Song Dynasty)
Decorative Features	Carved Lines	Direct painting on white body	Multiple techniques integrated
Manufacturing Process	Carved decoration in white glaze	Underglaze slip painting	Carved decoration and Underglaze slip painting

Note. All images are drawn from the dataset compiled for this study.

Yaozhou kilns were significant for northern celadon during the Song dynasty. The colour of the Yaozhou kilns is an olive green, a dark colour that is associated with the character of northern people. Yaozhou kiln porcelains are characterised by their decorations, predominantly crafted with the knife-carving technique. The firing of the Yaozhou kiln commenced during the Five Dynasties, continued throughout the Northern Song Dynasty, Jin, and persisted until the Yuan Dynasty. Nevertheless, the overall process of its fabrication appears to be in a state of decline. The Yaozhou kilns of the Five Dynasties were all carved in relief, with the entire decoration raised. The Northern Song Dynasty is characterised by the intaglio carving [62].

Folk Jun kilns are characterised by their relative simplicity in terms of the firing process when compared to other kilns. The Yaozhou kilns, in particular, require the application of a knife for carving, a process that demands a high level of skill and expertise. Jun kilns are characterised by a reduced level of complexity. The range of kiln firing is very wide. A variety of Jun kilns were discovered, each exhibiting distinctive body characteristics. These porcelain bodies ranged in colour from iron black to dark

grey, brownish-yellow, and even white, showcasing a rich diversity of body types. As previously stated, the milky glaze utilised in Jun kilns possesses a notable covering power. Consequently, Jun kilns exhibit minimal requirements for the body. The Jun kiln was a widely produced artefact, but the quality of these kilns was not guaranteed, and since the Song, Jin and Yuan dynasties, there has been a general decline in the quality of Jun kilns. The evolution of Jun porcelain can be characterised by a transition in colour from red to blue, accompanied by a shift in texture from delicate to rough [62].

The Folk Ding kiln has been observed to exhibit a range of variations, including the Black Ding, Purple Ding and Green Ding. These variations have been noted to feature carved floral decorations and imitations of the Cizhou kilns. It is also important to note that Purple Ding is a brown colour [62].

Jingdezhen has a unique condition of having kaolin clay. It was subsequently determined that kaolin was the optimum raw material for firing porcelain on a global scale. Following the Northern Song Dynasty, Jingdezhen porcelain is renowned for its bluish-white porcelain. This distinctive style is positioned between celadon and white porcelain, exhibiting a distinctive blue hue. The superior quality of porcelain produced in Jingdezhen was a consequence of the region's abundant supply of high-quality clay. The porcelain's whiteness is of a notably high quality [62].

The Longquan kilns are distinguished by their distinctive green hue in the production of celadon in the southern region. The optimal era for Longquan celadon is the Southern Song Dynasty. In the Northern Song Dynasty, Longquan was known for its production of pea green. However, under the Southern Song Dynasty, Longquan began to transform into bluish-green. The colour was notable for its evenness and moisture. The porcelain body of the Northern Song period is characterised by a thick exterior with a thin glaze and crackles, while the Southern Song body is distinguished by a thin exterior with a thick glaze, as shown in Table 3.9. This is the process of its progress [62].

In the southern region, there is a significant presence of black porcelain, which is also referred to as the Jian kiln. The Jian kilns were primarily responsible for the production of tea bowls. The term "partridge patch" is employed to denote a specific type of decoration. The texture of the tea bowls has been crafted to emulate the feathers of the partridge. Bionic decorations of Tea bowls have gained significant popularity in Jian kilns, including the rabbit hair striation, which exhibits a texture reminiscent of rabbit hair. In the domain of Jian kiln varieties, the initial category pertains to rabbit hair striation. The second is the oil drop pattern, which is characterised by a glaze that evokes the appearance of a drop of oil. In addition, a variety of spotting patterns are observed, including tortoiseshell and partridge patches [62].

Table 3.9. Comparative Analysis of Longquan Celadon Characteristics Across Song Dynasties

Characteristics	Northern Song Longquan Celadon	Southern Song Longquan Celadon
Image		
Porcelain Title	Mallet-shaped vase with animal-shaped handles in celadon glaze, Longquan kiln(Song Dynasty)	Mallet-shaped Vase in celadon glaze, Longquan kiln (Song Dynasty)
Glaze Color	Pea green	Bluish-green
Body and Glaze Characteristics	Thick body, Thin glaze	Thin body, Thick glaze
Surface Features	With crackles	Without crackles

Note. All images are drawn from the dataset compiled for this study.

The Jizhou kiln is near the Cizhou Kiln, and both share common characteristics in terms of their respective lifestyles. The distinctive features of Jizhou kilns include black glazed porcelain and a wide range of decorative techniques, such as leaf patterns, paper cut and tortoiseshell motifs, as described in Table 3.10. The Jizhou kiln has also been found to be engaged in the mass production of tea bowls [62].

Table 3.10. Decorative Techniques of Song Jizhou Kiln Black Glazed Porcelain

Decoration Technique	Leaf Pattern	Paper Cut Motif	Tortoiseshell Motif
Image			
Porcelain Title	Bowl with leaf decoration in black glaze, Jizhou kiln (Song Dynasty)	Bowl with applied plum blossom paper cut decoration in black glaze, Jizhou kiln (Song Dynasty)	Jar with tortoiseshell decoration in black glaze, Jizhou kiln (Song Dynasty)
Decorative Features	Yellow wood leaf pattern	Paper cut patterns	Black base with yellow-brown spots, similar to natural tortoiseshell
Common Vessel Types	Mainly tea bowls	Mainly tea bowls	Jars, vases, bowls

Note. All images are drawn from the dataset compiled for this study.

In the history of Chinese porcelains, the porcelain production process of the Yuan dynasty occupies an extremely important position. The primary kiln systems of the Yuan dynasty, including the Jun, Cizhou, Longquan and Dehua kiln systems, were constructed on the foundations laid by the Song dynasty. The advent of Yuan blue-and-white porcelain led to the establishment of Jingdezhen’s status as the porcelain capital [96].

During the Yuan Dynasty, the Jun kiln continued to produce traditional varieties, including sky-blue glaze, moon-white glaze, and bluish-green glaze with red spots. In comparison to the Song Jun Kiln, the Yuan Jun Kiln is distinguished by a coarse body, a multitude of pinholes on the glaze surface, a diminished lustre, and glaze colours that blend sky blue and moon white. The glaze does not extend to the base, and the interior and exterior of the foot are unglazed. The decoration of Jun kilns from the Yuan Dynasty was achieved through two distinct methods. The initial method was

the red spot decoration technique, in which copper-containing glaze was applied in an irregular shape to the surface of the vessel. The second technique was the piled decoration technique, which was unique to Yuan Jun kilns and absent from Song Jun kilns [20], as described in Table 3.11.

Table 3.11. Comparative Analysis of Yuan Jun Kiln Decoration Techniques

Decoration Technique	Red Spot Decoration	Piled Decoration
Image		
Porcelain Title	Bowl in bluish-green glaze with purple spots, Jun kiln(Yuan Dynasty)	Lotus petal-shaped Planter-stand in bluish-green and grape-purple glaze with piled decoration, Jun kiln(Yuan Dynasty)
Manufacturing Process	Irregular application of copper-containing glaze on vessel surface	Three-dimensional piled technique

Note. All images are drawn from the dataset compiled for this study.

The Cizhou Kiln of the Yuan Dynasty continued to produce porcelains based on the techniques of the Song Dynasty. During the Yuan Dynasty, the Cizhou Kiln was engaged in the production of porcelains with underglaze black over white ground, the main forms including jars, bowls and pillows, all with substantial size and rounded shape. The decorative techniques employed in the underglaze black over white ground porcelains from the Yuan Dynasty's Cizhou Kiln primarily consist of two methods. One method involves the application of a brown glaze over a white ground, accompanied by underglaze black. Another method involves the application of a bluish-green glaze over underglaze black [62], as shown in Table 3.12.

There are two principal features of Jingdezhen kilns of the Yuan Dynasty. Firstly, the successful production of Yuan blue-and-white porcelain is to be considered. The term "blue-and-white porcelain" is used to denote the application of cobalt pigment to a porcelain body, followed by the application of a transparent glaze, and firing at high temperatures to produce underglaze blue patterns. Yuan blue-and-white porcelain occupies a pivotal position in the history of Chinese porcelain production. Secondly, the successful production of underglaze red porcelains. Underglaze red is defined as

Table 3.12. Characteristics of Yuan Cizhou Kiln Porcelain

Characteristic	Description
Image	
Porcelain Title	Washer in bluish-green glaze over underglaze black, Cizhou kiln(Yuan Dynasty)
Decorative Technique	Bluish-green glaze over underglaze black

Note. All images are drawn from the dataset compiled for this study.

the application of copper pigment to the porcelain body, followed by the application of a transparent glaze, and firing at high temperatures, as reported in Table 3.13. It is evident that underglaze red and blue-and-white porcelain share certain similarities in terms of body, glaze, and firing techniques. However, it is important to note that these two types of porcelain also exhibit distinct differences. Firstly, underglaze red patterns tend to be simpler and less intricate in comparison to those observed in blue-and-white porcelain. Secondly, the subject matter depicted in underglaze red is less diverse than that seen in blue-and-white porcelain [62].

Traditional morphological authentication methods

The traditional morphological authentication of porcelains is based on three key elements: colour, form(shape) and pattern [66]. Porcelain surfaces are typically coated with one or two layers of glaze. The majority of glazes exhibit colouration, with even white glazes and transparent glazes demonstrating some degree of colouration, such as celadon or ivory white. Pure, colourless, transparent glazes are not present in archaeological porcelains. The pigmentation of glazes is determined by the colouring agents employed in their composition. During the Song and Yuan dynasties, black glaze, blue glaze, brown glaze, and purple glaze were all produced using iron oxide as the colouring agent, with variations occurring only in the proportion of iron content and the firing conditions. Green glaze and red glaze both utilise copper oxide as the colouring agent, while blue glaze employs cobalt oxide [66]. The classification of glaze colours in Chinese porcelains presented here is based on the dataset compiled for this study, as presented in Figure

Table 3.13. Comparative Analysis of Yuan Dynasty Jingdezhen Porcelain

Characteristics	Blue-and-White Porcelain	Underglaze Red Porcelain
Image		
Porcelain Title	Octagonal prunus vase with wave and dragon decoration in white glaze over underglaze blue, Jingdezhen kiln(Yuan Dynasty)	Yuhuchun vase with white incised pony decoration in greenish-white glaze over underglaze red, Jingdezhen kiln(Yuan Dynasty)
Colorant	Cobalt pigment	Copper pigment
Decorative Color	Underglaze blue patterns	Underglaze red patterns
Manufacturing Process	Cobalt application on body, transparent glaze, high-temperature firing	Copper application on body, transparent glaze, high-temperature firing
Pattern Complexity	Complex and intricate patterns	Simpler and less intricate patterns

Note. All images are drawn from the dataset compiled for this study.

3.2. These glazes can be systematically organised into several categories according to colour tone and decorative technique. White glazes include variations such as white, ivory white, bluish white, and moon white. Green glazes are further diversified into yellowish green, green, greyish green, celadon, greenish yellow, and bluish green. Purple tones, though less common, are distinguished as rose purple and grape purple. Blue is notably represented by sky blue. In terms of decorative methods, underglaze techniques include combinations such as white glaze with underglaze cobalt blue, and bluish green glaze with underglaze iron black. Other notable colours include brown and black glazes. This classification provides a foundational framework for analyzing porcelain colours.

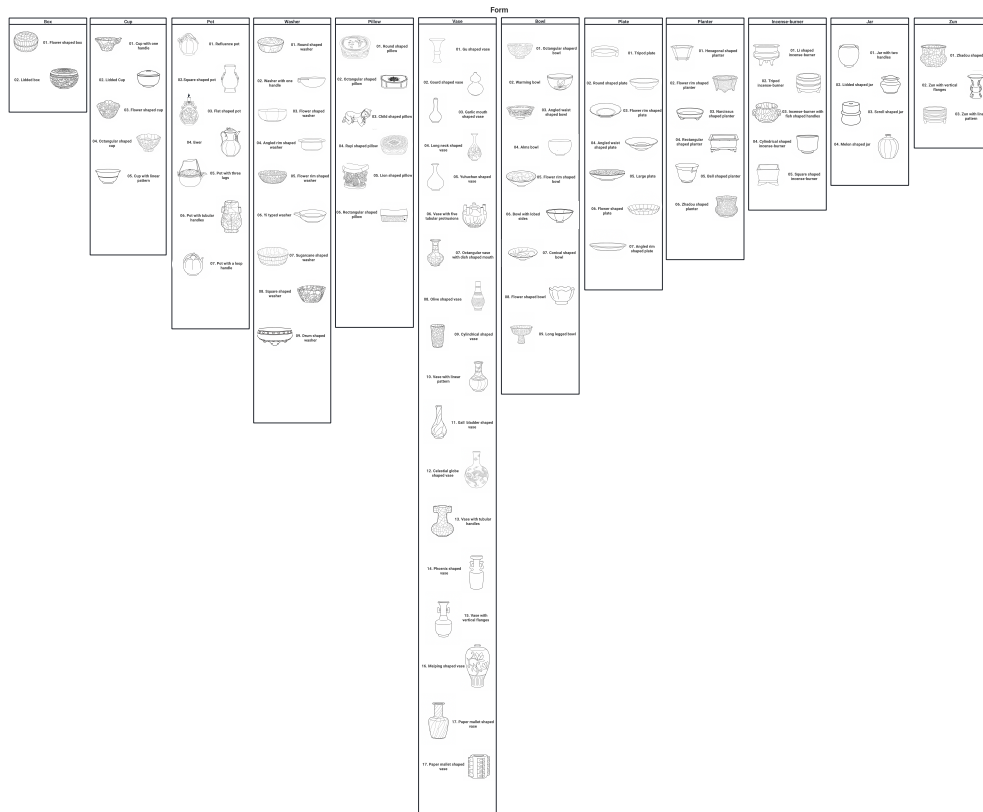


Note. All images are drawn from the dataset compiled for this study.

Figure 3.2. The Chinese porcelain glaze colour categories from the dataset

Porcelain's practical application and aesthetic appeal principally dictate the type of porcelain. Furthermore, it evolves by temporal shifts and advancements in porcelain production techniques. The typological classification of porcelain forms presented in this dataset categorises porcelain objects into major functional groups such as washers, pots, cups, boxes, pillows, vases, bowls, plates, planters, incense-burners, jars, and zuns, each further subdivided into specific shapes. For instance, vases appear in highly varied forms, including gu-shaped, gourd-shaped, Yuhuchun-shaped, Meiping-shaped, and phoenix-shaped types. Washers include forms like round, square, flower-shaped, and sugarcane-shaped, indicating ornamental as well as ritual use. Cups, bowls, and plates are diversified by rim profiles and base structures, such as lobed sides, angled

waists, and flower rims, suggesting shifts in design preference across dynastic periods. Incense-burners and planters also exhibit rich variations, including fish-shaped handles or narcissus-shaped bodies, as described in Figure 3.3.



Note. All images are drawn from the dataset compiled for this study.

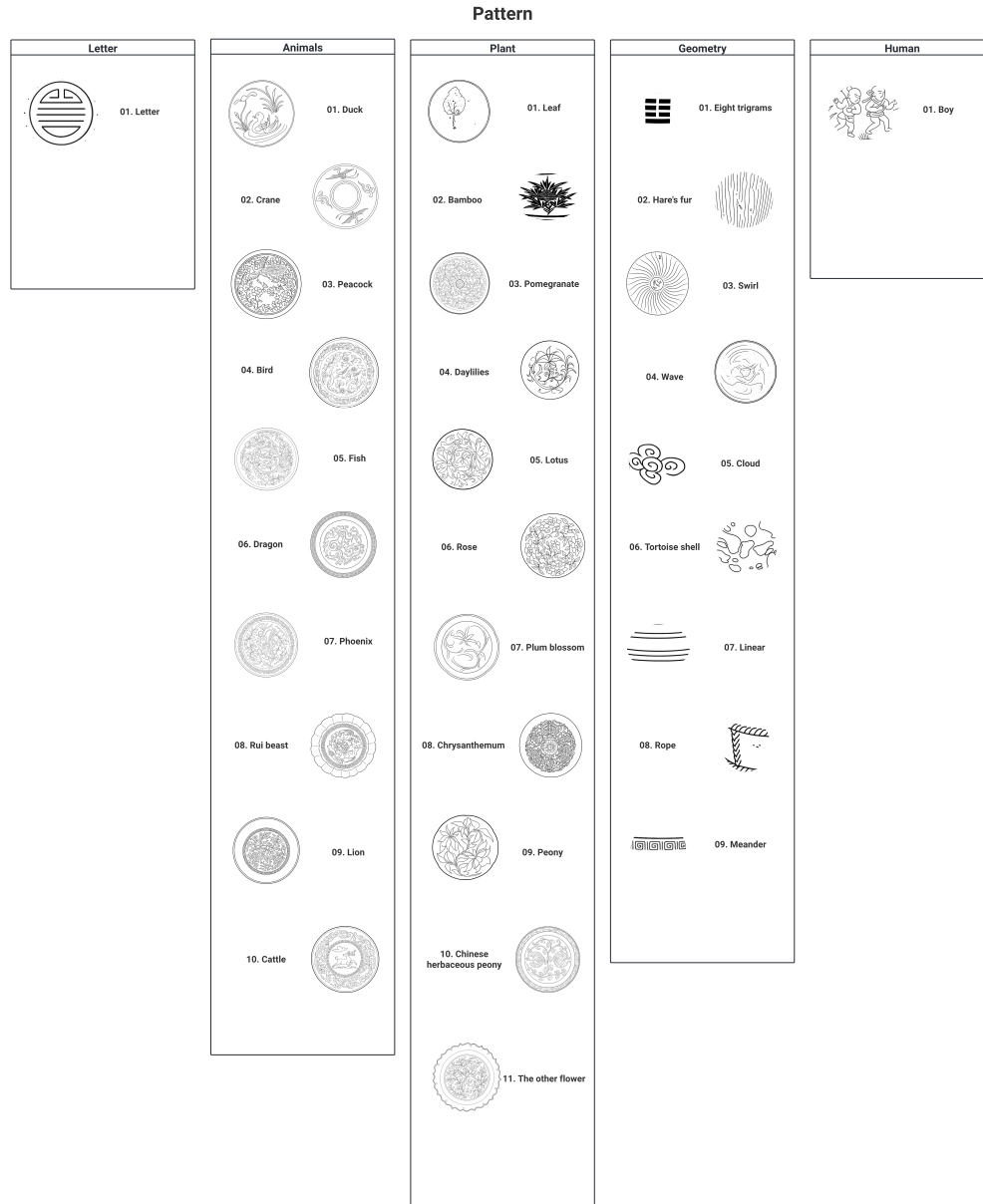
Figure 3.3. The Chinese porcelain form categories from the dataset

Decorative patterns constitute a vital aspect of porcelain morphology, particularly from the Yuan dynasty onwards. The classification of patterns reflects four production techniques, such as printed, incised, carved and painted. As shown in Figure 3.4, this figure presents a systematic taxonomy of decoration patterns found on Chinese porcelains, organised into five major categories, such as letter, animals, plant, geometry and human, each further subdivided into specific patterns. For instance, plants include leaves, bamboo, daylilies, lotus and rose. Animals include ducks, cranes, peacocks, fish and dragons. While the taxonomy presented in Figure 3.4 demonstrates 31 distinct decorative patterns across Song and Yuan porcelains, it is crucial to note that decorated pieces represent a minority within these periods. In Song dynasty porcelains, surface ornamentation was the exception rather than the rule. For instance, Ding kilns occasion-

ally featured impressed, incised, or carved motifs, while Yuan blue-and-white porcelains pioneered underglaze cobalt painting. However, the predominant aesthetic philosophy of both periods favored monochrome glazes without surface decoration, celebrating the purity of form and the subtle beauty of glaze effects themselves.

Statistical analysis of our dataset confirms this historical reality: only 24% ($n=1,746$) of the 7,263 porcelain images contain identifiable decorative patterns, with the remaining 76% consisting of undecorated monochrome porcelains. Among the decorated subset, the distribution across the 31 pattern categories exhibits extreme imbalance, with common motifs like lotus and peony appearing in hundreds of examples while rare patterns such as the rui beast or letter motifs occur in fewer than ten instances. This severe class imbalance, compounded by the overall scarcity of decorated porcelains, presented significant methodological challenges that ultimately led to the exclusion of pattern classification from our multi-task framework. The decision to omit pattern recognition as a fifth task was further justified by technical considerations. Pattern identification requires significantly higher image resolution than our other classification tasks, as fine details in carved or painted decoration may be imperceptible at standard imaging resolutions. Additionally, many patterns appear in partial or combined forms, creating ambiguous classification scenarios that would require multi-label rather than single-label classification architectures.

Future iterations of this research framework, particularly those incorporating Ming and Qing dynasty porcelains where decorated porcelains constitute the majority rather than the minority, could meaningfully integrate pattern classification as an additional task branch. The rich decorative traditions of these later periods, with their elaborate painted scenes, complex symbolic programs, and standardized pattern vocabularies, would provide sufficient training examples across pattern categories to support robust multi-task learning. Such temporal expansion would transform pattern decoration from a data-scarce edge case to a central classification objective, aligning the computational framework with the archaeological reality of Chinese porcelain development.



Note. All images are drawn from the dataset compiled for this study.

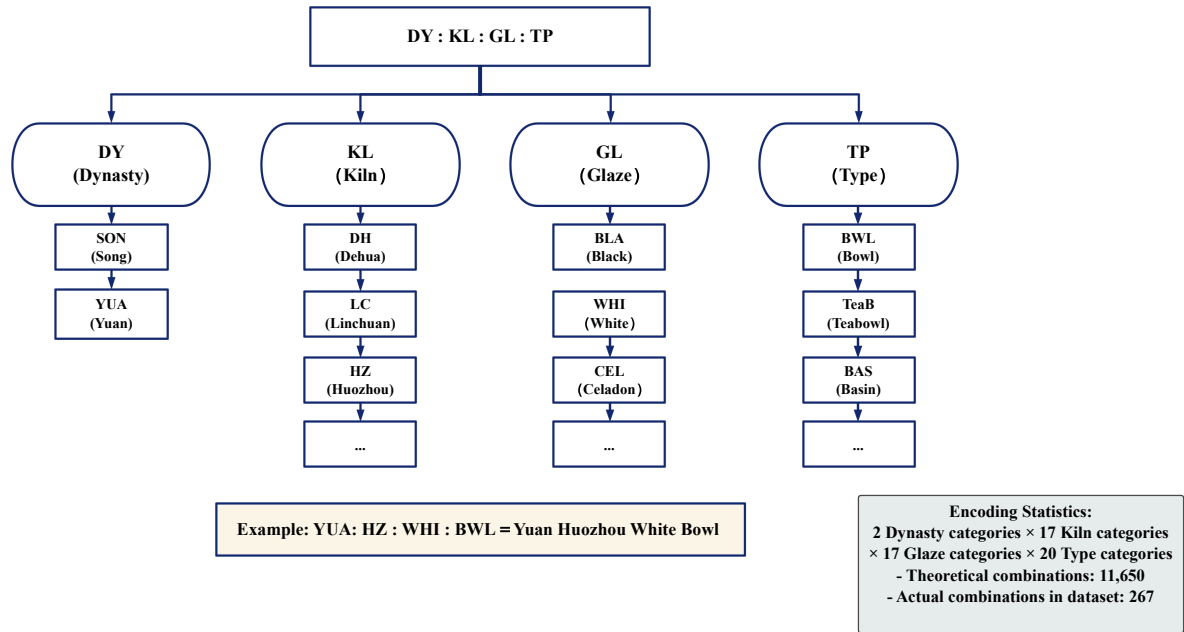
Figure 3.4. The Chinese porcelain pattern categories from the dataset

3.2.3 Hierarchical Label System and Historical Context

According to the traditional morphological authentication of porcelain, the image dataset is annotated with a four-axis hierarchical code:

DY : KL : GL : TP

where DY denotes the political period (Dynasty), KL(kiln) denotes the producing kiln, GL (Glaze) denotes the glaze colours, and TP (Type) denotes the functional porcelain type. The hierarchical label system shown in Figure 3.5 encodes each porcelain artifact using four attributes (Dynasty, Kiln, Glaze, Type), resulting in 267 observed combinations out of 11,650 theoretical possibilities. This sparse representation (2.3%) reflects the historical and technological constraints of porcelain production.



Note. Each attribute uses a 2-3 letter mnemonic code for dataset operations and unambiguous identification.

Figure 3.5. Hierarchical Label System for Porcelain Classification

The code performs a dual function: it provides an unambiguous primary key for dataset operations, and it embeds the principal variables that condition stylistic and technological change in Chinese porcelains. Each attribute is represented by a stable two- or three-letter mnemonic, reported in Table 3.14. The annotation process utilised the Labelbee tool, labelling images with four corresponding tags and saving them in CSV format.

Table 3.14. Hierarchical Label System

Axis	Tag	Typical values	$N_{\text{categories}}$
DY	SON, YUA	Song, Yuan	2
KL	RU, LQ, JZ, DH...	Ru, Longquan, Jingdezhen, Dehua...	17
GL	CEL, GRE, WHI...	Celadon, Green, White...	17
TP	BWL, VAS, JAR...	Bowl, Vase, Jar...	20

3.2.4 Train–Validation–Test Partitioning

The multi-attribute nature and severe combination imbalance of the porcelain dataset necessitated a carefully designed data splitting strategy that ensures meaningful representation across all subsets while preserving the integrity of rare combinations. Traditional random splitting would risk placing entire rare combinations exclusively in one subset, making reliable evaluation impossible.

We implemented an adaptive splitting algorithm that adjusts the partitioning strategy based on combination frequency. For singleton combinations ($n=1$), samples were assigned exclusively to the training set to maximise learning opportunity for these extremely rare cases. Doublet combinations ($n=2$) were deterministically split between validation and test sets, ensuring evaluation capability while acknowledging the absence of training examples. Small combinations ($3 \leq n < 10$) followed a 70-15-15 split approximation, with careful rounding to ensure at least one sample in each evaluation set. Standard combinations ($n \geq 10$) received the target 70-20-10 split, providing robust representation across all subsets.

This adaptive approach resulted in the following distribution: training set with 5,068 images, validation set with 1,392 images, and test set with 801 images. The slight discrepancy from the original 7,263 images suggests two samples were excluded during processing. The splitting algorithm successfully handled 4 singleton combinations (assigned exclusively to training), 6 doublet combinations (split between validation and test), 36 small combinations (3-9 samples), and 56 standard combinations (≥ 10 samples). Figure 3.6 shows the dataset splitting strategy based on combination size.

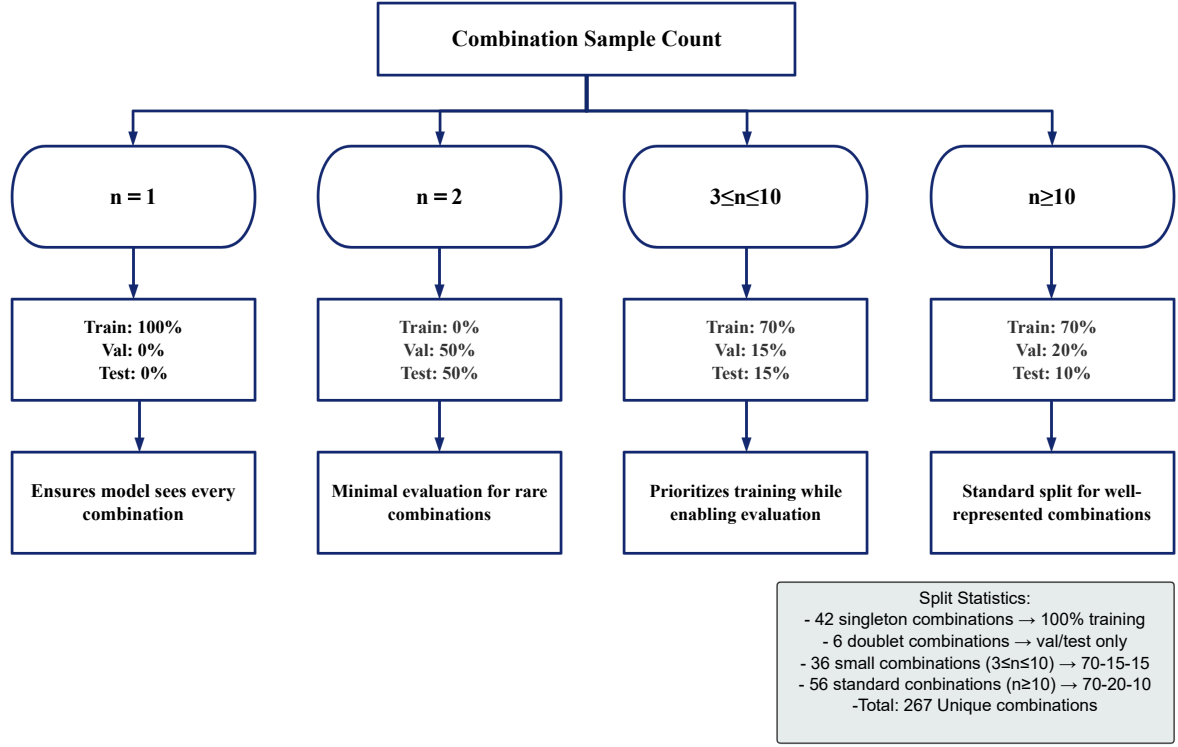


Figure 3.6. Adaptive Dataset Splitting Strategy Based on Combination Size

Post-splitting analysis reveals the effectiveness of this strategy: only 4 singleton combinations (3.9% of all combinations) appear exclusively in the training set, preserving their learning opportunity while minimising evaluation gaps. Remarkably, 67 combinations (65.7%) appear in all three subsets, indicating robust representation for the majority of classes. While 29 combinations (28.4%) lack validation representation due to the doublet assignment strategy, only 4 combinations (3.9%) lack test representation, ensuring comprehensive evaluation capability. This distribution demonstrates that despite extreme class imbalance, the adaptive splitting strategy successfully maintains evaluation integrity while maximising training opportunities for rare classes.

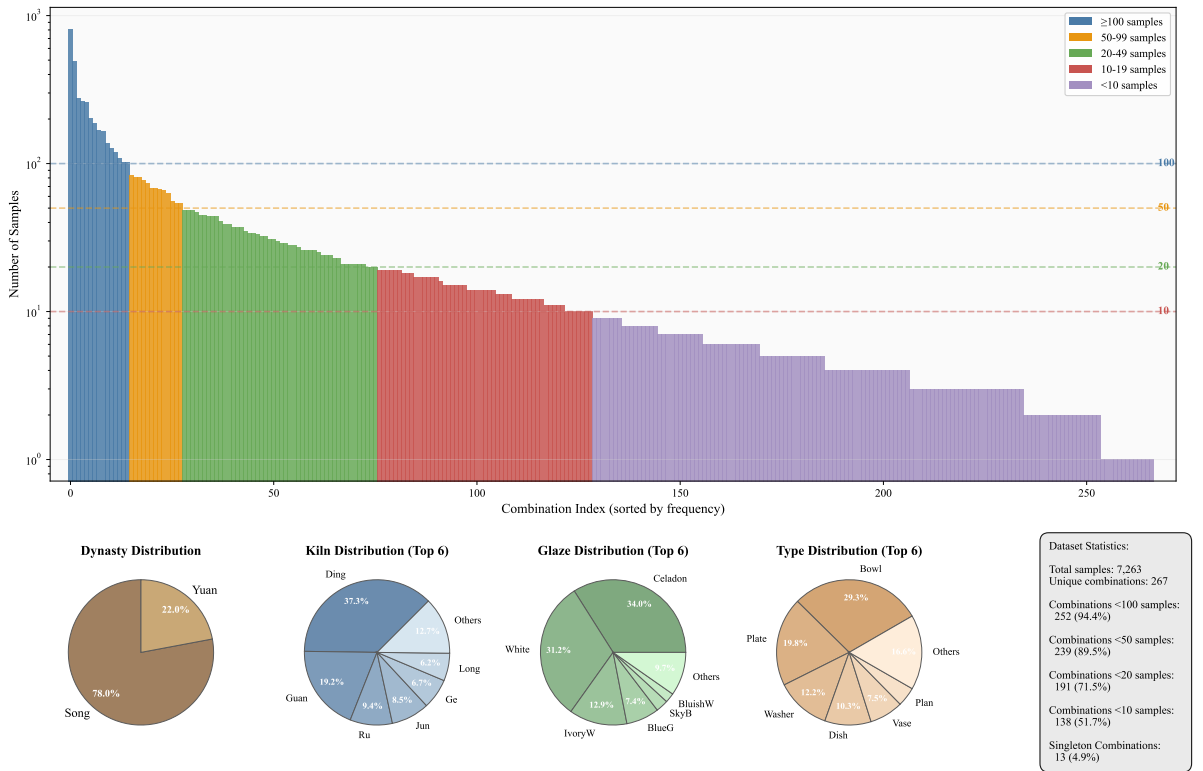
3.2.5 Quantitative Combination-Imbalance Analysis

The dataset exhibits severe combination imbalance that reflects differential preservation rates of archaeological Chinese porcelain. Analysis of the 7,263 original images reveals a highly inclined distribution across the multi-attribute classification space, as described in Table 3.15.

Table 3.15. Dataset Distribution Statistics

Total original images	7263
Unique combinations	267
Combinations with < 100 samples	252
Combinations with <50 samples	239
Combinations with <20 samples	191
Combinations with <10 samples	124

This extreme imbalance, with over 94% of combinations having fewer than 100 samples, poses fundamental challenges for deep learning approaches. The distribution follows a long-tail pattern where a small number of common porcelain types dominate the dataset, while the majority of unique combinations are severely underrepresented, as reported in Figure 3.7.

**Figure 3.7.** Long-tail Distribution of Porcelain Combinations

Decomposing the multi-attribute space into individual tasks reveals the hierarchical

nature of the imbalance, as reported in Table 3.16.

Table 3.16. Individual Task Statistics

Task	Number of Categories	Distribution Characteristics
DY	2	Binary classification (Song: 5,662, Yuan: 1,601)
KL	17	Dominated by Ding kiln (6,159 samples, 84.8%)
GL	17	White glaze comprises 68.0% of all samplese
TP	20	Bowls alone account for 73.2% of samples

The theoretical combination space of 11,560 possible combinations ($2 \times 17 \times 17 \times 20$) is sparsely populated, with only 267 combinations (2.3%) represented in the dataset. This extreme sparsity indicates that many historically plausible combinations are entirely absent, reflecting both actual production patterns and collection biases. For instance, while purple glaze appears in only 0.3% of samples and Peng kiln in 1.0%, their combination is virtually non-existent.

To address these challenges, we developed a comprehensive framework for mitigating imbalances. First, we applied threshold-based data augmentation to all class combinations below 50 samples, using carefully calibrated transformations: random horizontal flipping ($p = 0.5$), rotation ($\pm 30^\circ$), conservative colour jittering (brightness=0.2, contrast=0.2, saturation=0.2, hue=0.05), and random resized cropping (scale=0.8-1.0). This selective augmentation expanded the training set from 7,263 to 25,877 samples while maintaining visual authenticity. Second, we employed hierarchical sampling that operates at both task and combination levels, ensuring mini-batch diversity while over-sampling rare combinations with probability proportional to $1/\sqrt{n}$. Third, we adopted effective number weighting [13] with $\beta = 0.999$, providing smoother weight transitions than inverse frequency weighting while capping weights at 10.0 to prevent instability. These strategies collectively address the extreme class imbalance challenge, providing a foundation for effective multi-attribute porcelain classification across both common and rare combinations.

To address the severe combination imbalance, we employ effective number weighting that smoothly adjusts sample weights based on combination frequency. The effective number for a combination with n samples is calculated as (3.1):

$$E_n = \frac{1 - \beta^n}{1 - \beta} \quad (3.1)$$

The weight for combination i is then defined as the inverse of its effective number

(3.2):

$$w_i = \frac{1}{E_{n_i}} \quad (3.2)$$

$\beta = 0.999$ and n_i - is the number of samples in combination i .

During training, we implement hierarchical sampling to ensure adequate representation of rare combinations. The probability of sampling from combination c is proportional to (3.3):

$$P(c) \propto \frac{1}{\sqrt{n_c}} \quad (3.3)$$

n_c - is the number of samples in combination c .

Given the extreme imbalance with many singleton combinations, we adopt an adaptive splitting strategy based on combination frequency. The allocation ratios for training, validation, and test sets are determined by equation (3.4):

$$\text{Split}(n) = \begin{cases} (1.0, 0, 0) & \text{if } n = 1 \\ (0, 0.5, 0.5) & \text{if } n = 2 \\ (0.7, 0.15, 0.15) & \text{if } 3 \leq n < 10 \\ (0.7, 0.2, 0.1) & \text{if } n \geq 10 \end{cases} \quad (3.4)$$

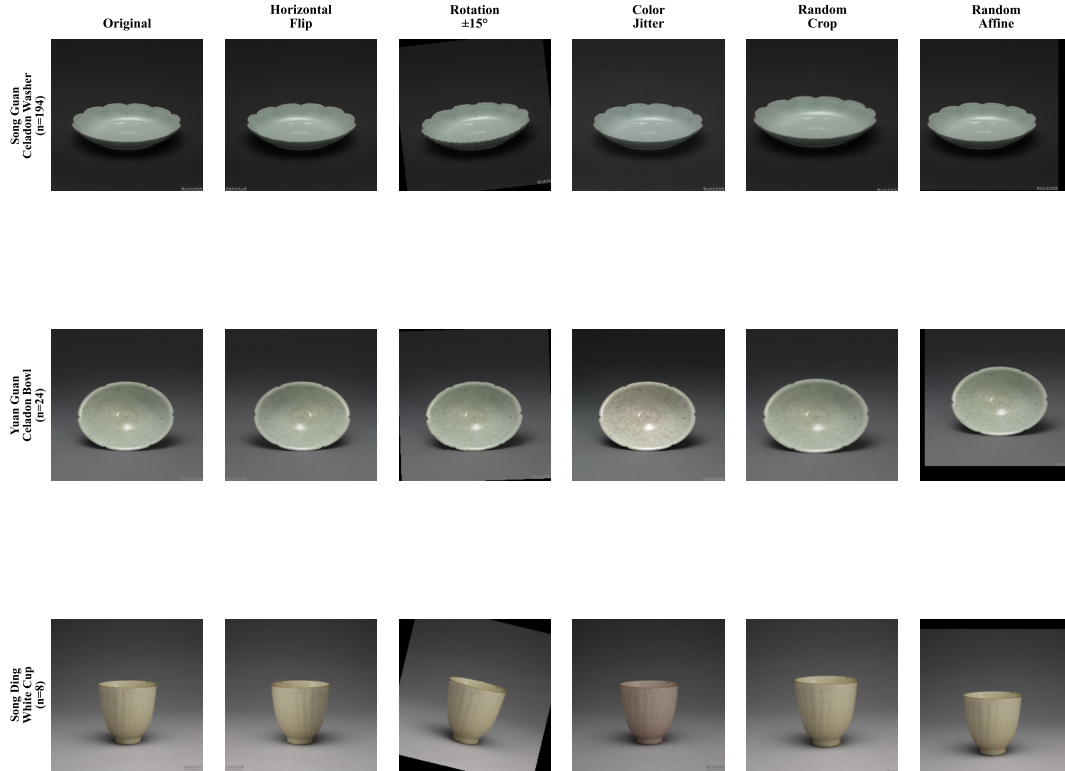
n - denotes the sample count for a combination and the triplet represents the (train, val, test) allocation ratios.

3.2.6 Image Pre-processing and Data-Augmentation

Our preprocessing pipeline addresses the dual challenges of standardizing heterogeneous museum photography conditions and augmenting rare combination representations without compromising the authenticity of porcelain visual characteristics.

For training images, we implemented a comprehensive augmentation pipeline that balances geometric and photometric transformations. Random resized cropping (scale 0.8-1.0) with 224×224 output simulates varying object distances while maintaining crucial detail visibility. Random horizontal flipping ($p = 0.5$) leverages the inherent symmetry of most porcelain forms, effectively doubling the training data without introducing artifacts. Rotation ($\pm 15^\circ$) and affine translation ($\pm 10^\circ$) account for photography angle variations common in museum documentation. Color jittering parameters (brightness=0.2, contrast=0.2, saturation=0.2, hue=0.1) were carefully calibrated through empirical testing to encompass realistic lighting variations while preserving the subtle color distinctions critical for glaze classification—excessive color augmentation would

render glazes like "white" and "ivory white" indistinguishable. Figure 3.8 illustrates the five augmentation techniques (horizontal flip, rotation, colour jitter, random crop, and affine transformation) applied to porcelain samples selected from different frequency ranges in the dataset.



Note: Samples selected from high-frequency ($n > 50$), mid-frequency ($10 < n \leq 50$), and low-frequency ($n \leq 10$) combinations

Figure 3.8. Data Augmentation Effects on Representative Porcelain Samples

Validation and test images undergo minimal preprocessing to ensure consistent evaluation. Images are first resized to 256×256 pixels, then center-cropped to the final size of 224×224 pixels, providing a slight zoom that eliminates potential border artifacts while maintaining the object's central features. This deterministic preprocessing ensures reproducible evaluation while accommodating the natural variation in object positioning within museum photographs. Both pipelines conclude with ImageNet normalization (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]), enabling effective transfer learning from pretrained models despite the domain shift from natural images

to cultural artifacts. This normalization strategy proved crucial for achieving rapid convergence during training.

Beyond online augmentation during training, we implemented offline targeted augmentation for severely underrepresented combinations as detailed in Section 3.2.5. This two-stage augmentation strategy, aggressive offline augmentation for combination balancing followed by moderate online augmentation for robustness enabled effective learning from rare combinations while maintaining the visual integrity essential for fine-grained porcelain analysis. The combination of thoughtful preprocessing and hierarchical augmentation strategies proved instrumental in achieving robust multi-task classification performance across the severely imbalanced dataset.

3.3 CNN Architecture and Multi-Task Learning Frameworks

3.3.1 Custom CNN Architecture

We design a custom convolutional neural network from scratch to establish a baseline performance without leveraging pre-trained knowledge. This baseline architecture serves a critical methodological purpose: it allows us to quantify the specific contributions of transfer learning by providing a direct comparison between training from random initialization versus utilizing pre-trained features. Without this baseline, it would be impossible to determine whether performance improvements stem from architectural choices, pre-trained knowledge, or the multi-task learning framework itself. Figure 3.9 shows the architectural comparison between baseline and transfer learning approaches: (Left) Custom CNN baseline with random initialisation, containing 4.7M parameters distributed between the convolutional backbone (99.4%) and task-specific heads (0.6%). (Right) Transfer learning models utilising ImageNet pre-trained weights, with model capacities ranging from 5.0M (MobileNetV3) to 45.6M (ResNet-101) parameters. Both approaches employ identical multi-task heads for fair comparison. The architectural similarity between Custom CNN and MobileNetV3 (4.7M vs 5.0M parameters) enables controlled evaluation of pre-training effects.

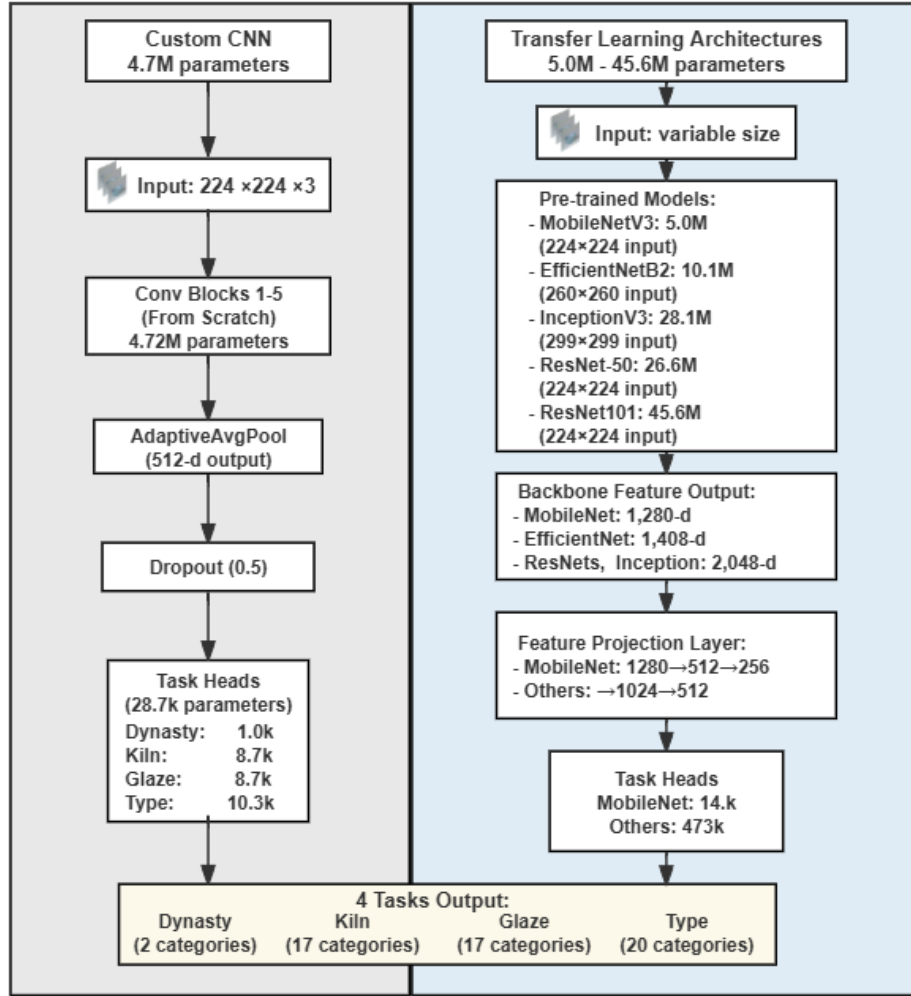


Figure 3.9. Architectural comparison between baseline and transfer learning approaches

The architecture adopts a straightforward design inspired by classical convolutional networks, particularly VGG [97]. It consists of five convolutional blocks with progressively increasing channel dimensions (32, 64, 128, 256, 512), where each block contains two 3×3 convolutional layers followed by batch normalisation [40], ReLU activation, and 2×2 max pooling. This design pattern has proven effective for image classification tasks while remaining computationally efficient. The final convolutional block outputs $7 \times 7 \times 512$ feature maps, which are reduced to a 512-dimensional vector through adaptive average pooling. This global pooling operation eliminates the need for predetermined input sizes while significantly reducing the parameter count compared to traditional fully connected layers.

For the multi-task configuration, the 512-dimensional features pass through a dropout layer ($p=0.5$) [40] before connecting directly to four task-specific linear classifiers: dynasty ($512 \rightarrow 2$), kiln ($512 \rightarrow 17$), glaze ($512 \rightarrow 17$), and type ($512 \rightarrow 20$). This direct connection, without intermediate shared layers, represents a minimal multi-task architecture that serves as a lower bound for multi-task performance. The complete model contains approximately 4.7 million trainable parameters, with the task-specific classification heads accounting for only 28,728 parameters (0.6%). This distribution underscores that the model’s capacity is almost entirely devoted to feature extraction rather than task-specific classification. The multi-task framework influences the learned representations through diverse supervisory signals. The four tasks—dynasty, kiln, glaze, and type—require different visual features, forcing the shared backbone to learn more comprehensive and robust representations than any single task would demand. For instance, type classification may rely primarily on varied forms and shapes, while glaze identification requires attention to surface texture and colour variations.

Notably, with 4.7M parameters, our custom CNN has a similar model size to MobileNetV3 (5.0M parameters), making their performance comparison particularly meaningful for isolating the impact of pre-trained features versus random initialization. This architectural similarity eliminates model capacity as a confounding factor, allowing us to attribute performance differences primarily to the quality of initial feature representations.

3.3.2 Multi-Task Learning Framework

Multi-task learning (MTL) aims to improve generalisation by leveraging domain-specific information contained in related tasks [89]. In the context of porcelain artefact analysis, the four attributes (dynasty, kiln, glaze, and type) exhibit natural correlations that make them suitable candidates for joint learning. These correlations arise from historical production constraints: specific kilns operated during certain dynasties, particular glazing techniques were exclusive to certain production sites, and vessel types often followed period-specific aesthetic preferences. By learning these tasks jointly, we hypothesise that the model can exploit these inherent relationships to improve overall classification performance.

We adopt the hard parameter sharing paradigm, the most commonly used approach in neural multi-task learning [89]. In this architecture, hidden layers are shared among all tasks while maintaining separate task-specific output layers, as described in Figure 3.10. This design offers several advantages: (1) it reduces the risk of overfitting by forcing the model to learn representations that generalise across tasks, (2) it improves computational

and storage efficiency by sharing the majority of parameters, and (3) it enables implicit knowledge transfer between related tasks through shared representations.

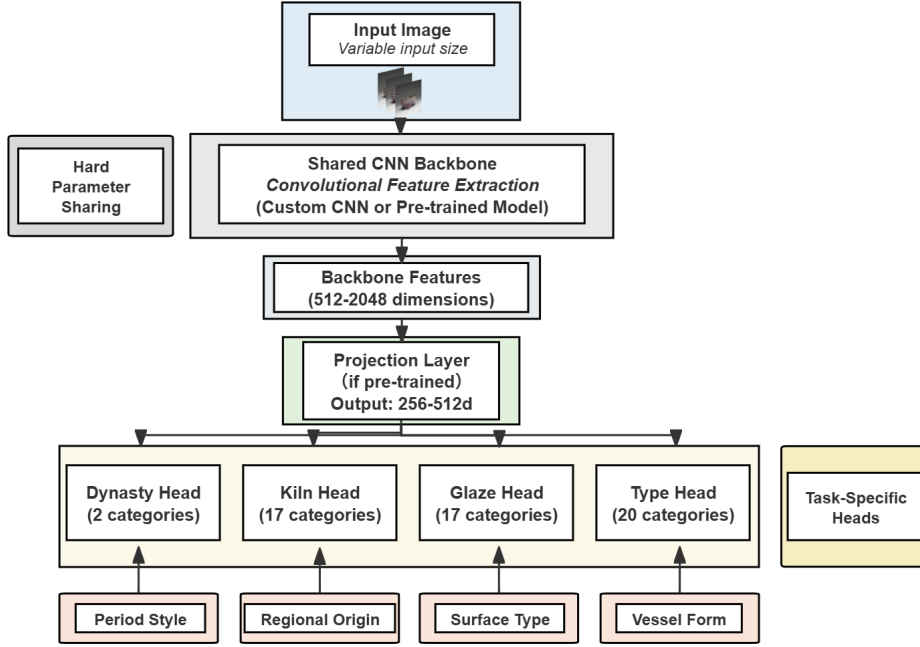


Figure 3.10. Multi-Task Learning Framework Conceptual Diagram

The shared backbone, whether custom-designed or pre-trained, processes input images to extract visual features that must simultaneously support four different classification objectives. This constraint encourages the emergence of more robust and generalizable features compared to single-task learning, where the representation might overfit to task-specific patterns. The effectiveness of this approach, however, depends critically on the relatedness of the tasks and the quality of the shared representation. We investigate these factors through our comparative experiments.

Following the shared feature extraction, we implement four parallel classification heads, each tailored to its specific task, as shown in Table 3.17.

- **Dynasty Classification Head:** A binary classifier distinguishing between Song (960-1279 CE) and Yuan (1271-1368 CE) dynasties. Despite being the simplest task with only two categories, dynasty classification requires the model to recognise subtle stylistic evolution and period-specific characteristics.
- **Kiln Classification Head:** A 17-way classifier identifying production sites.

This task demands recognition of regional characteristics, including local material properties, firing techniques, and decoration methods.

- **Glaze Classification Head:** Another 17-way classifier categorizing glaze types. Notably, despite having the same number of categories as kiln classification, glaze recognition proves significantly more challenging, requiring fine-grained discrimination of surface textures, colour variations, and reflective properties that may appear similar under varying photographic conditions.
- **Type Classification Head:** A 20-way classifier for vessel types. This task combines geometric understanding of three-dimensional shapes from two-dimensional projections with knowledge of functional categories that evolved throughout Chinese porelain history.

Table 3.17. Task Characteristics and Challenges

Task	Categories	Visual Features Required	Main Challenges
Dynasty	2	Period styles	Subtle evolution
Kiln	17	Regional traits, decoration methods	Site similarities
Glaze	17	Surface texture, color, reflectance	Photographic variance
Type	20	3D form, functional elements	View dependence

Each classification head consists of a single linear transformation from the shared feature space (512 dimensions) to the task-specific output space. The minimal design of these heads, comprising only 28,728 parameters total (0.6% of model capacity), which reflects our hypothesis that the primary challenge lies in learning appropriate visual representations rather than in the classification decision itself.

The multi-task framework creates a unique learning dynamic where gradients from four different objectives flow through the shared backbone during training. This can lead to both positive and negative transfer effects [108]:

- **Positive transfer** occurs when tasks provide complementary information. For instance, learning to recognize kiln sites might help with dynasty classification, as certain kilns were only active during specific periods. Similarly, glaze identification could benefit from kiln recognition, as certain glazing techniques were traditionally applied to specific kilns.
- **Negative transfer** emerges when task objectives conflict, causing the shared representation to compromise between incompatible feature requirements. This is particularly problematic when training from random initialisation, as the model

lacks a strong prior on what constitutes useful visual features. For instance, features optimal for distinguishing surface textures (crucial for glaze classification) might interfere with features needed for shape recognition (essential for type classification). The shared backbone must navigate these potentially conflicting requirements, which can lead to suboptimal representations for individual tasks. This theoretical challenge motivates our comparative study between random initialisation and pre-trained models, as pre-trained features might provide a more robust starting point for multi-task adaptation.

Multi-task learning offers particular advantages for our archaeological dataset of 25,877 images. Each training image provides supervision for four different objectives, potentially offering richer gradient signals during backpropagation. This multi-faceted supervision encourages the learning of more comprehensive features that capture various aspects of porcelain artefacts. This is especially valuable for rare categories that might have limited representation in the dataset. For example, a rare glaze type with few examples still benefits from the shared representations learned for more common glazes, as well as from correlated features learned for dynasty, kiln, and type classification.

Furthermore, the multi-task framework acts as a regularizer by preventing the model from overfitting to any single task. The shared parameters must satisfy multiple objectives simultaneously, encouraging the learning of more general features that are less likely to capture dataset-specific artefacts or biases.

3.3.3 Transfer Learning Models

Transfer learning has revolutionized computer vision by enabling the reuse of features learned from large-scale datasets for specialized tasks with limited data [43]. For our porcelain classification task, we leverage models pre-trained on ImageNet, which contains over 14 million images across 1,000 categories. While ImageNet primarily consists of natural images that differ substantially from archaeological porcelains, the low and mid-level features learned by these models—edges, textures, shapes, and patterns—remain highly relevant for porcelain analysis. We select five architectures representing different design philosophies and computational trade-offs in modern deep learning, as shown in Figure 3.11:

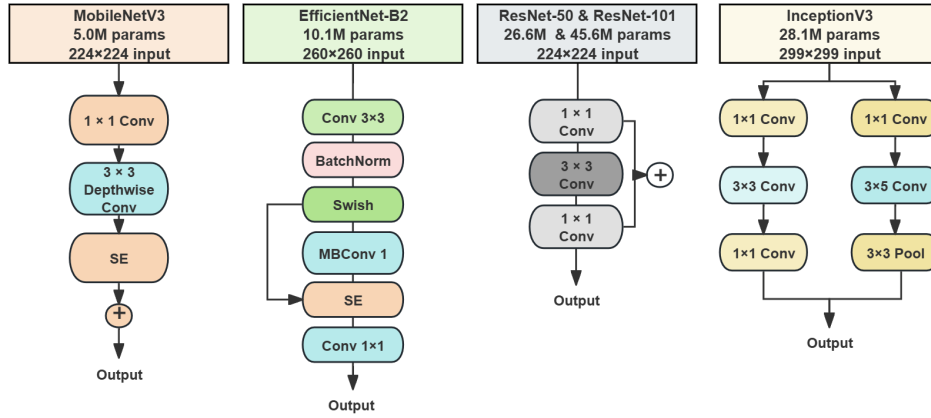


Figure 3.11. Architecture Comparison: Key Building Blocks

- **MobileNetV3-Large** [37]: This architecture epitomises efficiency through depth-wise separable convolutions, which decompose standard convolutions into depth-wise and pointwise operations. With only 5.0M parameters, MobileNetV3 incorporates squeeze-and-excitation modules and neural architecture search (NAS) optimisations. Its lightweight design makes it particularly attractive for deployment in resource-constrained museum environments while potentially offering sufficient capacity for porcelain classification.
- **EfficientNet-B2** [104]: Developed through systematic neural architecture search, EfficientNet employs compound scaling—simultaneously scaling network depth, width, and resolution according to principled coefficients. The B2 variant (10.1M parameters) represents a balanced point in the EfficientNet family, using 260×260 input resolution. This higher resolution may benefit fine-grained porcelain features such as glaze textures and decorative details.
- **ResNet-50 and ResNet-101** [32]: These architectures introduced residual connections that enable training of very deep networks by addressing the vanishing gradient problem. ResNet-50 (26.6M parameters) uses bottleneck blocks across 50 layers, while ResNet-101 (45.6M parameters) extends this to 101 layers. The residual connections have proven particularly effective for transfer learning, as they preserve gradient flow during fine-tuning.
- **InceptionV3** [101]: This architecture employs parallel convolutions with multiple kernel sizes (1×1 , 3×3 , 5×5) within "inception modules" to capture multi-scale features. With 28.1M parameters and requiring 299×299 inputs, InceptionV3

processes visual information at multiple granularities simultaneously. This multi-scale approach may prove advantageous for porcelain analysis, where both fine details (glaze texture) and global structure (vessel form) are diagnostic.

Table 3.18 presents the key specifications of five pre-trained models adapted for our multi-task framework. These architectures span a wide range of parameters (5.0M to 45.6M) and input resolutions 224×224 to 299×299 , representing diverse approaches to deep visual feature learning.

Table 3.18. Specifications of Pre-trained Models Adapted for Multi-task Porcelain Classification

Model	Params	Depth	Input Size	Feature Dim	Year	Key Innovation
MobileNetV3-L	5.0M	16	224^2	1,280	2019	Depthwise separable
EfficientNet-B2	10.1M	23	260^2	1,408	2019	Compound scaling
ResNet-50	26.6M	50	224^2	2,048	2015	Residual connections
InceptionV3	28.1M	42	299^2	2,048	2015	Multi-scale convolution
ResNet-101	45.6M	101	224^2	2,048	2015	Deeper residuals

Each pre-trained backbone was adapted to integrate seamlessly with our multi-task classification framework through a four-step process. First, the original classification head (trained for 1,000 ImageNet classes) was removed, and features were extracted from the penultimate layer: 1,280-dimensional vectors from MobileNetV3’s final pooling layer, 1,408-dimensional vectors from EfficientNet-B2’s top convolution, and 2,048-dimensional vectors from the average pooling layers of ResNet-50/101 and InceptionV3.

Second, these high-dimensional features were projected into task-appropriate representations via two-layer projection modules. For MobileNetV3, the 1,280-dimensional features were mapped through a $1,280 \rightarrow 512 \rightarrow 256$ pipeline with ReLU activations and dropout rates of 0.5 and 0.35, yielding 256-dimensional embeddings. For the other architectures, projection followed a $2,048 \rightarrow 1,024 \rightarrow 512$ (ResNet/Inception) or $1,408 \rightarrow 1,024 \rightarrow 512$ (EfficientNet) structure, with ReLU activations and 0.3 dropout after each layer.

Third, task-specific heads were applied to these projected features: MobileNetV3 employed direct linear mappings from the 256-dimensional features to the task logits, whereas the other models used two-layer heads ($512 \rightarrow \text{hidden} \rightarrow \text{classes}$), with hidden dimensions set to 128 for the binary dynasty task and 256 for the more complex kiln, glaze, and type tasks.

Finally, all models used architecture-specific preprocessing pipelines, including image resizing to 224×224 , 260×260 , or 299×299 , normalization to ImageNet statistics, and corresponding data augmentation strategies. For InceptionV3, the

auxiliary classifier branch was repurposed during training to produce predictions for all four tasks, contributing 0.3 of the total loss to provide additional gradient signals to earlier layers; this branch was disabled during inference.

We adopt a full fine-tuning strategy with differential learning rates, assigning a lower rate of 10^{-5} to the pre-trained backbone to preserve transferable features, and a higher rate of 10^{-3} to the projection layers and task-specific heads to enable rapid adaptation. This configuration mitigates catastrophic forgetting while allowing effective task-specific refinement. For MobileNetV3, the backbone learning rate is increased to 10^{-4} —ten times higher than for the other models—reflecting the architecture’s lightweight design and its greater tolerance for aggressive parameter updates, as confirmed in our implementation notes, as described in Table 3.19.

Table 3.19. Fine-tuning configurations by model and component

Model	Component	Learning Rate	Weight Decay	Dropout
MobileNetV3	Backbone	1e-4	1e-5	—
	Projection	1e-3	1e-5	0.5, 0.35
	Task heads	1e-3	1e-5	—
Other models	Backbone	1e-5	1e-4	—
	Projection	1e-3	1e-4	0.3
	Task heads	1e-3	1e-4	0.2

Common settings: gradient clipping = 1.0, optimizer = AdamW.

3.3.4 Loss Function and Optimization

Our multi-task learning objective combines individual task losses through weighted aggregation, allowing us to balance the contributions of tasks with varying difficulty levels and class distributions. The total loss function is formulated as (3.5):

$$\mathcal{L}_{total} = \sum_{t \in \mathcal{T}} \lambda_t \mathcal{L}_t \quad (3.5)$$

where $\mathcal{T} = \{\text{dynasty, kiln, glaze, type}\}$ represents the four tasks, λ_t denotes the task-specific weight, and $\mathcal{L}_t$ is the the individual task loss. Each task $\mathcal{L}_t$ employs weighted cross-entropy to address class imbalance (3.6):

$$\mathcal{L}_t = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^{C_t} w_{t,c} \cdot y_{i,c}^{(t)} \cdot \log(\hat{y}_{i,c}^{(t)}) \quad (3.6)$$

where N is the batch size, C_t is the number of Categories for task t , $w_{t,c}$ represents the class weight for class c in task t , $y_{i,c}^{(t)}$ is the ground truth label and $\hat{y}_{i,c}^{(t)}$ is the

predicted probability.

Based on empirical validation and task complexity analysis, we assign differential weights to each task:

- **Dynasty** ($\lambda_{\text{dynasty}} = 1.0$): Binary classification task serving as the baseline.
- **Kiln** ($\lambda_{\text{kiln}} = 1.2$): Moderately difficult, with 17 geographically distributed categories.
- **Glaze** ($\lambda_{\text{glaze}} = 2.0$): Assigned the highest weight due to subtle visual differences among 17 categories.
- **Type** ($\lambda_{\text{type}} = 1.5$): Focuses on complex shape recognition across 20 vessel categories.

These weights were determined through preliminary experiments, where we observed that glaze classification consistently underperformed despite having the same number of classes as kiln identification. The 2.0 weight for glaze classification effectively doubles its gradient contribution, encouraging the model to learn more discriminative features for surface texture analysis.



Figure 3.12. Loss configuration and class distribution analysis

Figure 3.12 shows the loss configuration and class distribution analysis. (a) Task-specific loss weights (λ) assigned based on task difficulty, with *glaze* receiving the highest weight ($\lambda = 2.0$) despite having the same number of classes as *kiln*. (b) Class distribution for the *glaze* task showing severe imbalance, with the most frequent class (Celadon) appearing 6,800 times while rare classes appear fewer than 1,000 times. The red line indicates the effective weights calculated using $\beta = 0.999$, which provide more balanced compensation than inverse frequency weighting. (c) Binary *dynasty* classification showing a moderate imbalance ratio of $1.33\times$ between Song (14,777 samples) and Yuan (11,100 samples) dynasties. (d) Summary statistics reveal that *glaze* exhibits the highest class imbalance ($34.1\times$) among all tasks, justifying its higher loss weight.

To address the severe combination imbalance, we employ effective number weighting that smoothly adjusts sample weights based on combination frequency [13]. This approach prevents extreme weights for very rare classes while still providing meaningful compensation for class imbalance. We further cap the maximum weight at 10.0 to prevent gradient instability.

We employ AdamW optimizer with decoupled weight decay [58], which has shown superior performance for fine-tuning pre-trained models. The optimization configuration uses differential learning rates across model components:

- **Backbone parameters:** learning rate = $1e-5$ (or $1e-4$ for MobileNetV3)
- **Other parameters (projection + heads):** learning rate = $1e-3$
- **Weight decay:** $1e-4$ (or $1e-5$ for MobileNetV3)

The $10\times$ higher learning rate used for MobileNetV3’s backbone reflects the need for more aggressive updates in efficient architectures. Lightweight models often require different optimisation dynamics compared to over-parameterised networks such as ResNet or Inception.

We implement cosine annealing [57] to smoothly decay the learning rate over training (3.7):

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos \left(\frac{t\pi}{T} \right) \right) \quad (3.7)$$

where η_t is the learning rate at epoch t , $\eta_{\max}$ is the initial learning rate, $\eta_{\min} = 10^{-6}$ is the minimum learning rate, and T is the total number of epochs. This schedule helps the model converge to better local minima by enabling larger updates in early training and progressively finer adjustments as convergence nears.

Table 3.20. Optimization settings for multi-task training

Component	Setting	Value
Optimizer	Type	AdamW
	β_1, β_2	0.9, 0.999
Learning Rates	Backbone (MobileNet)	1e-4
	Backbone (Others)	1e-5
	Projection + Heads	1e-3
Weight Decay	MobileNet	1e-5
	Others	1e-4
LR Schedule	Type	Cosine Annealing
	$T_{\max}$	50 epochs
	$\eta_{\min}$	1e-6
Gradient Clipping	Max norm	1.0
Batch Size	ResNet/Inception	16–24
	MobileNet	64
	EfficientNet	48
Task Weights (λ)	Dynasty/Kiln/Glaze/Type	1.0 / 1.2 / 2.0 / 1.5
Class Weight Strategy	Method	Effective Number
	β	0.999
	Max weight	10.0

To ensure stable training, we apply gradient clipping with a maximum norm of 1.0 across all parameters. This prevents gradient explosion, particularly important during early training stages when the randomly initialised task heads adapt to pre-trained features.

For InceptionV3 with auxiliary classifiers, we extend the loss computation to include auxiliary predictions (3.8):

$$L_{\text{total}} = L_{\text{main}} + \alpha L_{\text{aux}}, \quad (3.8)$$

where $\alpha = 0.3$ weights the auxiliary loss contribution. The auxiliary branch provides additional gradient signals at intermediate network depths, potentially improving feature learning in earlier layers. Each auxiliary task head follows the same class weighting strategy as the main heads.

3.3.5 Training Strategy

Our training protocol spans 50 epochs with early stopping based on validation performance. Each epoch processes the entire augmented training set (25,877 samples) through weighted random sampling, ensuring balanced exposure across rare and common categories. The training process monitors multiple indicators to ensure stable convergence and prevent overfitting. During training, we track both task-specific and aggregate metrics. The per-epoch training loop computes running averages of individual task losses and accuracies, providing immediate feedback on optimization progress. Gradient norm monitoring ensures stable updates, particularly crucial during the initial epochs when task-specific heads rapidly adapt to pre-trained features.

Given the multi-task nature and severe class imbalance in our dataset, we employ a comprehensive suite of evaluation metrics that capture different aspects of model performance.

- **Top-1 and Top-5 Accuracy** [91]: While top-1 accuracy provides the standard measure of exact predictions, top-5 accuracy acknowledges the inherent ambiguity in fine-grained porcelain classification. Many porcelain pieces exhibit transitional features or characteristics of multiple production sites. Top-5 accuracy considers a prediction correct if the true label appears within the model's five most confident predictions, particularly relevant for tasks with numerous visually similar categories: kiln (17 classes) where regional styles overlap, glaze (17 classes) where subtle color variations like "moon white" vs "ivory white" challenge even expert discrimination, and type (20 classes) where functional forms evolved gradually. The glaze task especially benefits from top-5 evaluation, as surface treatments often combine multiple techniques or represent transitional styles between established categories, making definitive single-label classification inherently ambiguous.
- **F1 Scores** [98]: We compute both macro and weighted F1 scores for each task. The macro F1 score, calculated as the unweighted mean across all classes, treats rare categories as equally important as common ones—crucial for archaeological applications where rare specimens often carry significant historical value. The weighted F1 score, weighted by class support, provides a more balanced view of overall performance. For model selection, we use the average macro F1 across all tasks (3.9):

$$F1_{\text{avg}} = \sum_{t \in T} \lambda_t \cdot F1_{\text{macro}}(t) \quad (3.9)$$

- **Precision and Recall**: Computed both per-class and macro-averaged, these

metrics reveal task-specific trade-offs. High precision is crucial for dynasty classification to avoid historical misattribution, while high recall matters more for rare glaze identification to ensure comprehensive cataloging. The macro-averaged versions provide unbiased estimates across all categories.

- **Confusion Matrix Analysis [82]:** Beyond aggregate metrics, confusion matrices reveal systematic error patterns that align with archaeological knowledge. We expect higher confusion between historically related categories (e.g., Longquan and Guan kilns, both producing celadon) than between distinct traditions (e.g., Ding white porcelain vs. Jian black porcelain).

Table 3.21. Evaluation metrics used for multi-task porcelain classification

Metric	Definition	Justification	Computation
Top-1 Accuracy	Proportion of correct predictions	Standard metric; intuitive but potentially misleading for imbalanced data	$\frac{\text{Correct predictions}}{\text{Total samples}}$
Top-5 Accuracy	True label appears in top-5 predictions	Captures ambiguity in fine-grained tasks (e.g., kiln, glaze, type)	$\frac{\text{True label in top-5}}{\text{Total samples}}$
F1-macro	Unweighted mean of per-class F1 scores	Treats all classes equally; crucial for rare, archaeologically important types	$\frac{1}{C} \sum_{i=1}^C \text{F1}_i$
F1-weighted	Class-support-weighted mean of F1 scores	Reflects real-world class imbalance in evaluation	$\sum_{i=1}^C \frac{n_i}{N} \cdot \text{F1}_i$
Precision (macro)	Average proportion of correct positive predictions	Indicates prediction reliability; important for documentation accuracy	$\frac{1}{C} \sum_{i=1}^C \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}$
Recall (macro)	Average proportion of correctly identified instances	Measures completeness; key for cataloging rare types	$\frac{1}{C} \sum_{i=1}^C \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}$
Confusion Matrix	Class-wise prediction distribution	Reveals systematic error patterns (e.g., between celadon-producing kilns)	Cell (i, j) = count of class i predicted as class j
<i>Notation:</i> C = number of classes; n_i = samples in class i ; N = total samples; TP, FP, FN = true/false positives/negatives			
Multi-task Aggregation: $\text{F1}_{\text{avg}} = \frac{1}{4} \left(\text{F1}_{\text{dynasty}}^{\text{macro}} + \text{F1}_{\text{kiln}}^{\text{macro}} + \text{F1}_{\text{glaze}}^{\text{macro}} + \text{F1}_{\text{type}}^{\text{macro}} \right)$			

Table 4.11 summarizes the comprehensive evaluation framework employed to assess multi-task porcelain classification performance. Each metric is selected to address specific challenges in our imbalanced dataset: while top-1 accuracy provides intuitive performance measures, top-5 accuracy acknowledges the inherent ambiguity in fine-grained porcelain classification where transitional features are common. The F1 scores, computed both as macro (unweighted) and weighted averages, ensure that rare porcelain types, often of greatest archaeological significance, receive appropriate consideration alongside common categories. Precision and recall metrics evaluate the reliability and completeness of classifications, while confusion matrices reveal systematic error patterns that can be interpreted through archaeological knowledge. For multi-task model selection, we aggregate performance using the average macro F1 score across

all four tasks, preventing optimization bias toward simpler tasks while encouraging balanced performance across the diverse classification challenges.

Validation occurs at the end of each epoch on the held-out validation set (1,392 images) without data augmentation. The validation loop computes all metrics in evaluation mode with dropout disabled and batch normalization using running statistics. This provides unbiased performance estimates for model selection. Model checkpoints are saved based on validation performance, specifically when the average macro F1 score improves. This composite metric prevents scenarios where models optimize for easy tasks (dynasty classification with 2 classes) while neglecting challenging ones (glaze classification with 17 classes and extreme imbalance). The patience mechanism triggers early stopping after 10 epochs without improvement, preventing overfitting while allowing sufficient exploration.

3.3.6 Experimental Set-up

This subsection details the experimental configuration used to evaluate multi-task learning approaches for porcelain classification. We describe the computational infrastructure, training procedures, evaluation protocols, and interpretability methods employed to ensure reproducible and comprehensive experiments.

All experiments were conducted on a single workstation equipped with an NVIDIA GeForce RTX 2070 SUPER GPU (8GB VRAM) and an AMD Ryzen 7 3700X processor—an 8-core, 16-thread CPU operating at a base frequency of 3.60 GHz (boost up to 4.4 GHz). The system included 16 GB of DDR4 RAM and a 1 TB NVMe SSD used for dataset storage and model checkpoints, ensuring high I/O throughput during training. The operating system was Windows 10 (64-bit).

The 8 GB GPU memory constraint played a significant role in experimental design, particularly influencing batch size configuration across architectures. While the RTX 2070 SUPER provides 2,560 CUDA cores and 9.1 TFLOPS of FP32 performance, memory limitations required careful tuning to balance utilization with training stability. The relatively modest 16 GB system RAM also motivated efficient data loading strategies to avoid CPU-side memory bottlenecks during preprocessing.

Despite these constraints, the setup proved sufficient to train models ranging from 4.7M parameters (Custom CNN) to 45.6M parameters (ResNet-101), enabled by adaptive batch sizing and memory-efficient training strategies.

The implementation is based on PyTorch 2.0.1 [78] with CUDA 11.8 support, providing improved performance and memory efficiency compared to earlier versions—particularly advantageous in our memory-constrained setup. Torchvision 0.15.2 is used for load-

ing pre-trained ImageNet models and standardized transformation pipelines, ensuring alignment with common computer vision practices.

Of particular note is PyTorch 2.0.1’s support for `torch.compile()`, a feature that enables ahead-of-time graph compilation for potential runtime acceleration. However, all experiments in this study were conducted in eager execution mode to preserve transparency during debugging. The use of CUDA 11.8 ensures compatibility with the RTX 2070 SUPER (Turing architecture) while enabling access to modern GPU optimizations, as shown in Table 3.22.

Table 3.22. Software stack used for model development and evaluation

Component	Version / Description
Core Framework	PyTorch 2.0.1+cu118 (CUDA 11.8)
Computer Vision	Torchvision 0.15.2+cu118
Model Zoo	TIMM 1.0.16
Data Processing	NumPy 1.24.4, Pandas 1.5.3, Pillow 9.4.0
Machine Learning	Scikit-learn 1.3.0 (metrics, preprocessing)
Visualization	Matplotlib 3.7.1, Seaborn 0.12.2
Development Environment	Python 3.11.4 (Anaconda)

3.4 Results and Analysis

This section presents a comprehensive evaluation of all model architectures on the test set, analysing both task-specific and overall multi-task performance. The evaluation encompasses 801 test images representing 141 unique porcelain combinations, providing robust evidence for model comparison.

3.4.1 Quantitative Performance Evaluation (RQ2.1,RQ2.2)

Table 3.23 presents the complete performance metrics across all six models and four classification tasks, revealing a substantial performance gap between transfer learning approaches and the custom CNN baseline. The custom CNN achieves an average F1-macro score of 0.430, while all pre-trained models exceed 0.733, with MobileNetV3 achieving 0.767 and ResNet-101 reaching 0.764, a 78.1% relative improvement that demonstrates the critical importance of transfer learning for fine-grained visual recognition tasks with limited domain-specific data.

This performance gain is particularly pronounced in challenging multi-class scenarios: for kiln classification (17 classes), the custom CNN achieves only 0.300 F1-macro

Table 3.23. Comprehensive Test Results on Test Set

Model	Task	Accuracy	Top-5 Acc	F1-macro	F1-weighted	Precision	Recall
Custom CNN	Dynasty	0.625	–	0.385	0.769	0.500	0.313
	Kiln	0.688	0.938	0.300	0.764	0.352	0.272
	Glaze	0.781	1.000	0.472	0.756	0.440	0.530
	Type	0.688	0.969	0.564	0.702	0.628	0.595
	<i>Average</i>	0.695	0.969	0.430	0.748	0.480	0.427
MobileNetV3	Dynasty	0.896	–	0.848	0.893	0.878	0.827
	Kiln	0.869	0.986	0.739	0.859	0.844	0.685
	Glaze	0.811	0.980	0.734	0.810	0.796	0.733
	Type	0.824	0.974	0.746	0.820	0.844	0.710
	<i>Average</i>	0.850	0.980	0.767	0.845	0.840	0.739
EfficientNet-B2	Dynasty	0.889	–	0.835	0.884	0.871	0.811
	Kiln	0.866	0.979	0.725	0.858	0.761	0.710
	Glaze	0.787	0.974	0.631	0.780	0.723	0.605
	Type	0.816	0.973	0.743	0.812	0.841	0.695
	<i>Average</i>	0.840	0.975	0.733	0.834	0.799	0.705
ResNet-50	Dynasty	0.905	–	0.864	0.903	0.882	0.850
	Kiln	0.859	0.983	0.736	0.852	0.819	0.703
	Glaze	0.769	0.990	0.646	0.766	0.756	0.610
	Type	0.799	0.970	0.740	0.796	0.812	0.698
	<i>Average</i>	0.833	0.981	0.746	0.829	0.817	0.715
ResNet-101	Dynasty	0.906	–	0.867	0.905	0.882	0.855
	Kiln	0.871	0.980	0.732	0.865	0.787	0.700
	Glaze	0.785	0.983	0.693	0.781	0.742	0.678
	Type	0.833	0.978	0.764	0.831	0.830	0.727
	<i>Average</i>	0.849	0.980	0.764	0.845	0.810	0.740
InceptionV3	Dynasty	0.889	–	0.843	0.887	0.855	0.832
	Kiln	0.850	0.984	0.716	0.844	0.782	0.682
	Glaze	0.700	0.985	0.673	0.704	0.738	0.681
	Type	0.793	0.969	0.730	0.791	0.800	0.683
	<i>Average</i>	0.808	0.979	0.740	0.806	0.794	0.719

Note: Top-5 accuracy not applicable for binary dynasty classification. Best performance per metric is shown in **bold**.

while MobileNetV3 reaches 0.739, a 146% improvement suggesting that ImageNet pre-training provides essential low-level visual features (edges, textures, shapes) that transfer effectively to porcelain surface analysis despite the substantial domain gap.

The differential benefits across tasks reveal interesting patterns about feature transferability: dynasty classification shows the highest absolute performance (0.906 accuracy with ResNet-101) and consistent $\sim 45\%$ improvement, benefiting from high-level semantic features; kiln attribution maintains robust performance (0.869–0.871 accuracy) despite having the most classes, indicating pre-trained features capture subtle texture and color variations; glaze classification exhibits the smallest relative improvement ($\sim 47\%$), suggesting more domain-specific features; and type classification benefits moderately ($\sim 20\%$ improvement), indicating partial overlap with ImageNet’s object recognition capabilities. The variation among pre-trained architectures (0.808–0.850 average accuracy) demon-

strates that architectural choices matter even within the transfer learning paradigm, with MobileNetV3's superior performance despite fewer parameters than ResNet-101 suggesting that recent innovations like squeeze-and-excitation blocks provide better feature extraction for this domain, offering practical advantages for resource-constrained deployment scenarios in field archaeology.

In conclusion, to answer **RQ2.2**, our results confirm that transfer learning significantly enhances the classification of porcelain artifacts across dynasty, glaze, kiln, and type tasks. The use of pre-trained CNNs not only accelerates convergence and stabilizes training but also improves classification accuracy and robustness, making it a highly effective approach for tasks with limited domain-specific training data.

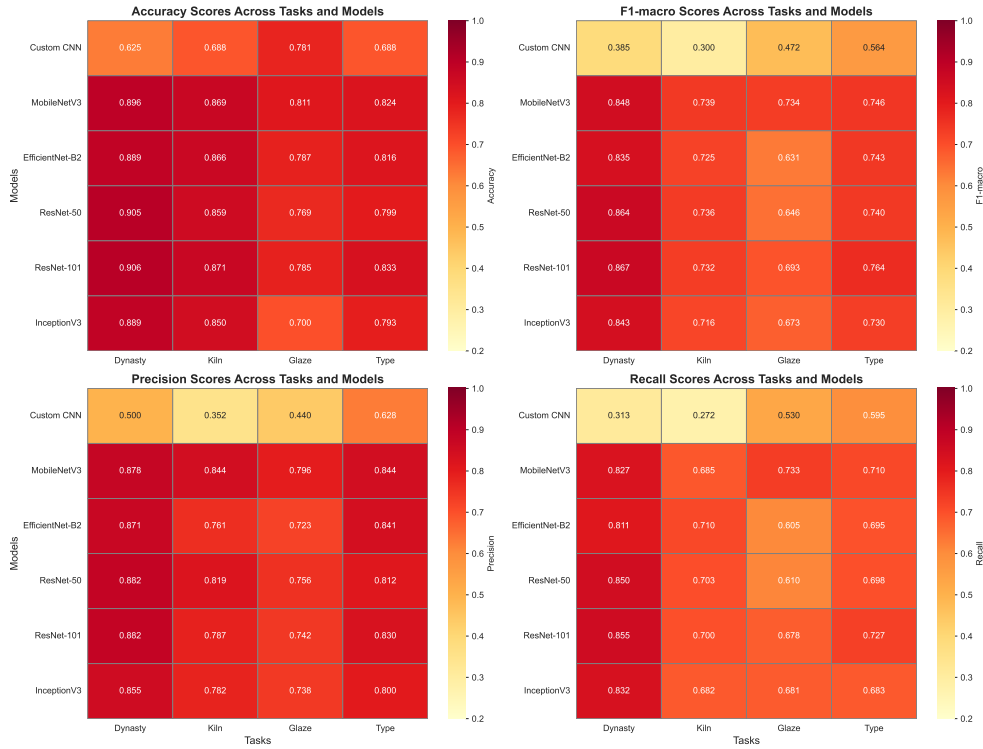


Figure 3.13. Comprehensive Performance Metrics Heatmap

Figure 3.13 visualizes the $F1$ -macro scores across all 24 model-task combinations, revealing consistent patterns in task difficulty and model performance. The average $F1$ -macro scores across all models for each task are: Dynasty (**0.813**), Type (**0.718**), Kiln (**0.693**), and Glaze (**0.659**), establishing a clear difficulty hierarchy. This ordering correlates with the number of categories (2, 20, 17, and 17, respectively) and the inherent visual complexity of each classification task.

MobileNetV3 demonstrates remarkable consistency across tasks, achieving the highest *F1-macro* score for Dynasty (0.848) and Glaze (0.734), and the second-highest for Kiln (0.739) and Type (0.746). This balanced performance is particularly valuable for multi-task learning applications, where no single task should dominate the optimization. In contrast, EfficientNet-B2 exhibits higher performance variance across tasks ($\sigma^2 = 0.0078$) compared to MobileNetV3 ($\sigma^2 = 0.0024$), suggesting less stable feature representations for diverse classification objectives.

The training time analysis reveals an unexpected finding: EfficientNet-B2 requires 1,128.7 minutes (18.8 hours) to train, 11.0 $\times$ longer than MobileNetV3 (102.7 minutes)—despite having only 2.02 $\times$ more parameters. This disproportionate increase stems from EfficientNet-B2’s higher input resolution (260 $\times$ 260 vs. 224 $\times$ 224) and compound scaling strategy, which increases computational complexity approximately quadratically with resolution. The extended training time poses practical challenges for hyperparameter tuning and iterative model development in real-world scenarios.

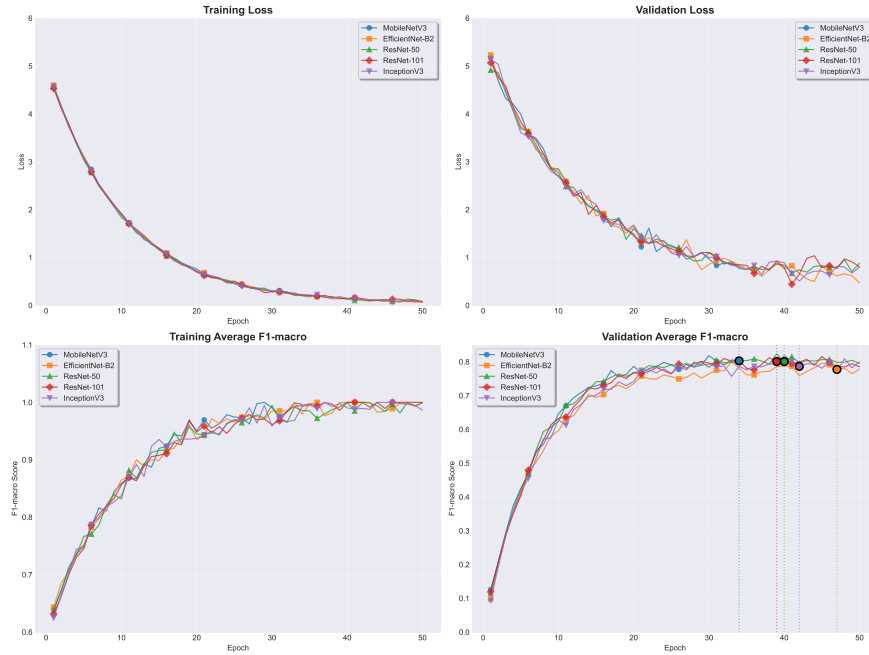


Figure 3.14. Training Dynamics of Transfer Learning Models

Figure 3.14 presents the training and validation curves for the five transfer learning models. MobileNetV3 exhibits rapid convergence, reaching its optimal validation F1 score of 0.811 at epoch 34, after which early stopping was triggered. The smooth convergence and minimal gap between training and validation curves indicate effective regularization without signs of underfitting. ResNet-50 and ResNet-101 show similar

convergence patterns, with their best validation performance observed at epochs 40 and 39, respectively. The marginal difference in final performance (F1: 0.808 vs. 0.804) suggests that increased network depth provides limited benefits in this setting, likely due to the relatively low spatial resolution of porcelain images compared to those in natural image datasets. InceptionV3 presents a unique case: initial training for 50 epochs yielded an F1 score of 0.790 at epoch 50. Extending training to 60 epochs improved performance to 0.796 at epoch 42, after which performance plateaued. This modest 0.76% improvement suggests that deeper and more complex architectures may require extended training when applied to fine-grained classification tasks. EfficientNet-B2 exhibits the highest volatility in its training curve, with validation loss fluctuating notably between epochs 40–50. Despite achieving a reasonable final performance (F1: 0.781), the unstable training dynamics, combined with prolonged training time, make it less favorable for practical deployment scenarios.

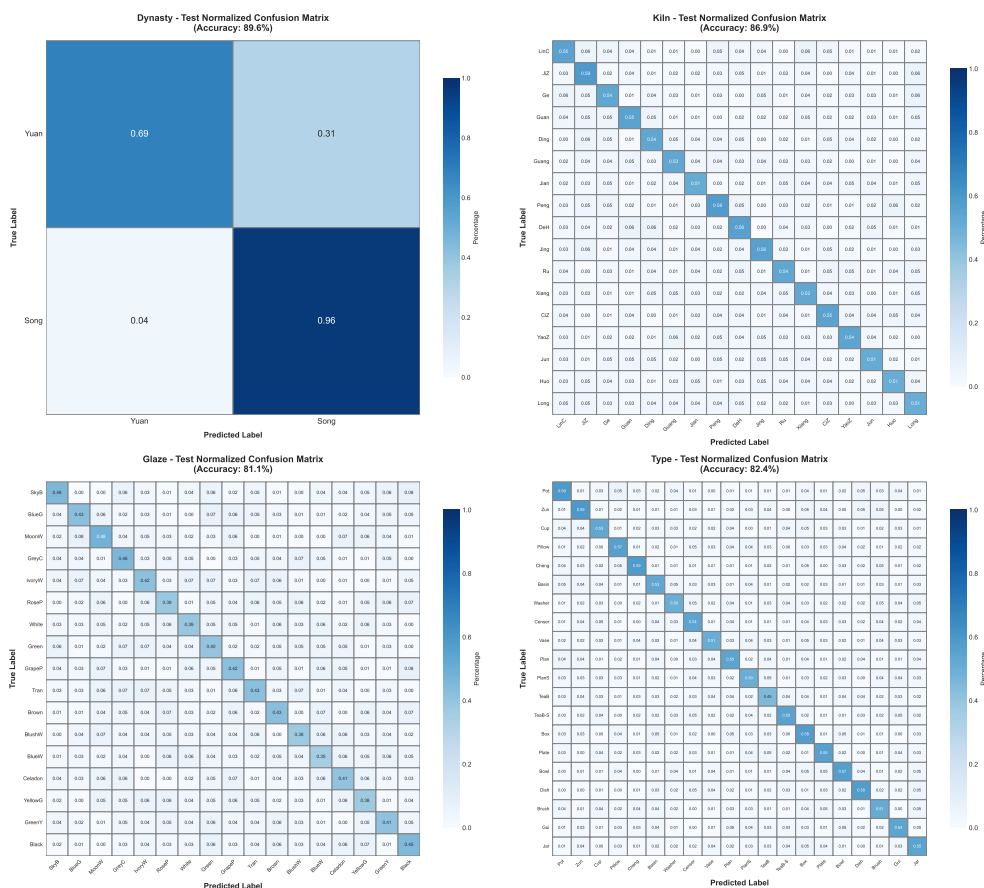


Figure 3.15. MobileNetV3 Multi-Task Confusion Matrices on Test Set

The confusion matrices for MobileNetV3 (Figure 3.15) reveal distinct, historically

grounded error patterns across the four classification tasks. In dynasty classification, the 2×2 matrix shows marked asymmetry, with 31% of *Yuan* pieces misclassified as *Song* compared to only 4% in the reverse direction—reflecting the documented continuity of Song production techniques into the Yuan period, particularly in Jun and Cizhou kilns. Kiln classification errors align with known geographical and stylistic relationships: the highest rates occur between *Guan* and *Ge* (11% and 8% mutual confusion), from *Linchuan* to *Jizhou* (17% and 33%), and in widespread misclassification of *Ding* toward multiple kilns, consistent with its historically influential white-bodied ware. In glaze classification, confusion is concentrated among chemically and visually similar types, most notably *Ivory White* to *White* (35%) and the reverse (16%), as well as frequent mutual errors among blue-tinted glazes such as SkyB (sky-blue), BlueG (bluish-green), and BlueW (bluish-white). Type classification shows predominantly morphological confusions: *Planter* and *Planter-stand* (14% and 11% mutual), *Bowl* to *Dish* (18%), and *Brush Holder* to *Vase* (50%)—the latter exacerbated by limited sample size. Together, these patterns demonstrate how visual similarity, historical continuity, and data scarcity interact to shape systematic classification errors.

The quantitative evaluation establishes **MobileNetV3** as the optimal architecture for multi-task porcelain classification, achieving the highest average F1-macro score (0.767) while maintaining exceptional parameter efficiency. However, these aggregate metrics mask important task-specific variations that warrant deeper investigation. While dynasty classification appears straightforward with accuracies exceeding 88%, the underlying error patterns reveal historically meaningful continuities. In contrast, the glaze classification task, with F1 scores ranging from 0.631 to 0.734 through five models, proves substantially more difficult and necessitates detailed analysis to uncover which specific glaze types pose the greatest challenges and why.

In response to the first research question **RQ2.1**, the results indicate that MobileNetV3 and ResNet101 emerge as the most robust models, demonstrating high accuracy and strong generalization across validation and test sets. EfficientNet-B2 and InceptionV3 also perform well but show minor overfitting. These findings suggest that MobileNetV3 is the most effective model for multi-task classification of porcelain artifacts, offering the best balance of accuracy and generalization. The following section provides comprehensive task-specific analysis, examining performance variations across categories, identifying systematic biases, and linking classification difficulties to the intrinsic properties of Chinese porcelain taxonomy.

3.4.2 Task-Specific Analysis (RQ2.1)

This section provides an in-depth analysis of MobileNetV3’s performance across the four classification tasks, examining the relationship between class characteristics, data distribution, and classification accuracy. The study reveals task-specific challenges and patterns that inform both the theoretical understanding of porcelain classification and practical deployment considerations. All training sample counts reported in this analysis refer to the augmented dataset (threshold-based augmentation), which reflects the actual training conditions of the model.

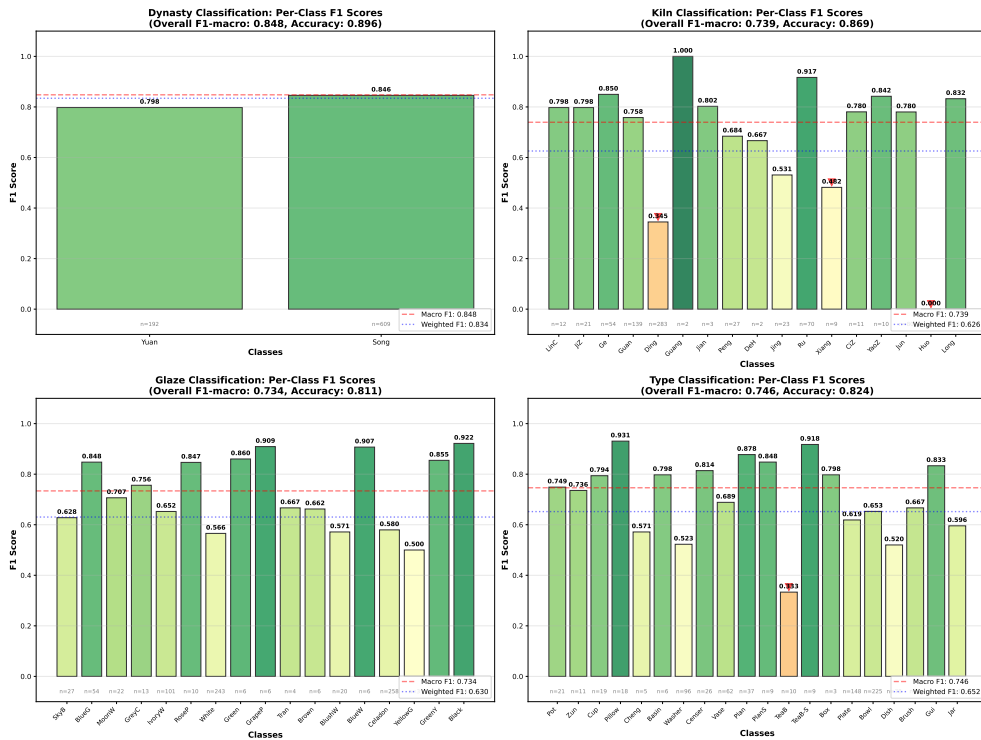


Figure 3.16. MobileNetV3: Per-Class F1 Performance Analysis

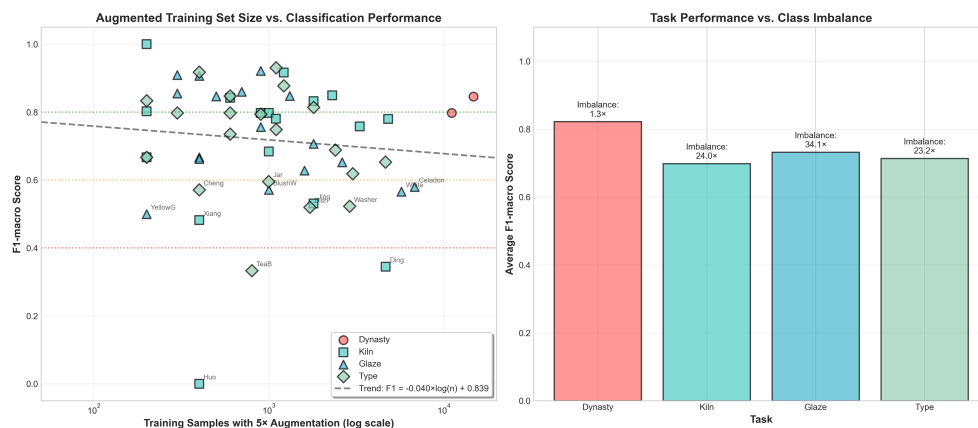
Dynasty classification achieves the highest overall performance (F1-macro: 0.848, Accuracy: 0.896) among all tasks, yet exhibits notable asymmetry in per-class performance (Figure 3.16). Yuan dynasty samples achieve only 0.798 F1 score compared to Song’s 0.846, despite having 11,100 after augmentation). This 5.7% performance gap reflects documented historical continuity, where Yuan (1271–1368 CE) deliberately maintained and evolved Song dynasty (960–1279 CE) porcelain traditions.

The confusion matrix analysis (Section 3.4.1) corroborates this pattern, showing 31% of Yuan pieces misclassified as Song, while only 4% of Song pieces are misclassified

as Yuan. This asymmetric confusion aligns with art historical evidence discussed in Section 3.2.3, where Yuan kilns, particularly Jun, Cizhou, Longquan, and Dehua, were explicitly built upon Song foundations. For instance, Yuan Jun kilns continued producing traditional sky-blue, moon-white, and bluish-green glazes with red spots, but with subtle differences: coarser porcelain bodies, more pinholes on glaze surfaces, and the introduction of piled decoration techniques absent in Song Jun wares.

The model's difficulty in distinguishing Yuan pieces suggests that it may be correctly recognizing genuine stylistic continuities rather than merely reflecting class imbalance or insufficient training data.

Kiln classification presents moderate difficulty (F1-macro: 0.739, Accuracy: 0.869) with substantial per-class performance variation. The 17 production sites demonstrate F1 scores ranging from 0.000 (Huo) to 1.000 (Guang), revealing complex patterns that reflect the hierarchical nature of Song porcelain production systems documented in Table 3.3. Despite modest training sizes, several kilns achieve very high recognition performance: Guang attains $F1 = 1.000$ with only 200 augmented samples; Ru reaches $F1 = 0.917$ with 1,217 augmented samples, likely aided by salient motifs such as “crab claw” patterns and “fish-scale crackles”; and Jun achieves $F1 = 0.842$ with 4,800 augmented samples, benefitting from its milky, opaque glaze. These results suggest that distinctive and stable stylistic cues can compensate for limited data volume. Conversely, some kilns remain difficult to classify even with comparatively ample data: Ding ($F1 = 0.345$ with 4,650 augmented samples), Jingdezhen ($F1 = 0.531$ with 1,800), and Huo/Huozhou ($F1 = 0.000$ with 400). Widespread imitation of Ding-style decorative and firing techniques across sites increases inter-class similarity and blurs boundaries, so additional data does not readily translate into robust discriminative signals. The scatter plot analysis reveals a weak negative correlation between training sample size and performance ($r = -0.040$), as shown in Figure 3.17. The poor performance of **Ding kiln** is particularly revealing, as the only white porcelain producer among the Five Great Kilns, its thin glaze enabled distinctive carved, incised, and printed decorations (Table 3.7). However, these decorative techniques were widely imitated across multiple production sites, creating substantial inter-class similarity. The “upside-down firing” technique adopted mid-Song Dynasty, resulting in unglazed rims, became a distinctive feature but was also copied by folk kilns, further complicating visual classification.



Note. Training samples include $5\times$ data augmentation.

Figure 3.17. Impact of Class Distribution on MobileNetV3 Performance

Glaze classification exhibits the lowest average performance (F1-macro: 0.734, Accuracy: 0.811) and the highest class imbalance ratio ($34.1\times$). The per-class analysis reveals systematic performance degradation for chemically and visually similar glaze types. Among glaze types, White (F1 = 0.566), Celadon (F1 = 0.580), Ivory White (F1 = 0.667), and Yellowish-Green (F1 = 0.500) are harder to distinguish. Their chemical variability and sensitivity to firing conditions (e.g., iron-oxide content, kiln atmosphere) yield broad visual ranges and high intra-class dispersion, diminishing the model’s ability to learn consistent cues. By contrast, Black (F1 = 0.922), Sky Blue (F1 = 0.848), and Grape Purple (F1 = 0.855) perform strongly. They typically possess narrower compositional ranges and striking visual signatures, e.g., Jian and Jizhou’s “rabbit-hair” and “oil-drop” effects for Black, the milky opacity of Jun’s Sky Blue, and the copper-oxide-based, vivid Grape Purple—making them reliably separable even in compressed photographs. As discussed in Section 3.2.2, glaze coloration is primarily influenced by the presence of metallic oxides and specific firing conditions. The poor performance of **White** and **Celadon** glazes can be attributed to their wide chemical variability. Both Song and Yuan celadons employed iron oxide (Fe_2O_3) as the dominant colorant; however, variations in iron content and kiln atmosphere produced a wide chromatic range from pale green to deep olive. Moreover, even within the same kiln, visual output varied significantly. For instance, **Longquan kiln** produced *pea green* celadon in the Northern Song period and *bluish-green* celadon during the Southern Song era by adjusting the body-to-glaze thickness ratio and modifying firing techniques (Table 3.9). These glazes are better classified due to their narrower chemical compositions and distinctive visual signatures that remain discernible even after photographic capture and compression. This suggests that visual salience and compositional stability contribute

significantly to classification accuracy.

Type classification achieves moderate performance (F1-macro: 0.746, Accuracy: 0.824) across 20 morphological categories. The analysis reveals systematic patterns linking form complexity and functional ambiguity to classification accuracy. For morphological types, the best results ($F1 > 0.80$) occur for Bowl ($F1 = 0.931$), Vase ($F1 = 0.878$), and Pillow ($F1 = 0.918$). These categories exhibit clear, consistent morphological boundaries (e.g., canonical *Meiping* and *Yuhuchun* shapes) and are well represented in the dataset, enabling the model to learn stable shape-based features. Lower-performing types ($F1 < 0.60$) include Tea Bowl ($F1 = 0.333$), Water Pot (Cheng; $F1 = 0.571$), Dish ($F1 = 0.520$), and Washer ($F1 = 0.523$). Functional overlap and morphological ambiguity drive confusion: tea bowls resemble regular bowls in common top-down imagery, and the Washer class aggregates diverse shapes (round, square, floral, etc.) by function rather than form, yielding low intra-class consistency and higher misclassification rates.

Classification errors in these low-performing categories stem from functional overlap and morphological ambiguity. For example, tea bowls, while often finished with characteristic black glazes and decorative patterns, share structural similarities with regular bowls, especially when captured from top-view angles. Likewise, the washer class demonstrates high intra-class diversity, ranging from round to floral shapes, which undermines morphological consistency. This functional but visually diverse grouping poses a significant challenge for appearance-based classifiers.

The comparative analysis reveals performance patterns that are consistent with established archaeological understanding. First, the observed performance ordering (*Dynasty* > *Type* > *Kiln* > *Glaze*) inversely correlates with the number of classes, yet directly reflects archaeological distinctiveness: binary dynasty classification benefits from clear temporal boundaries, whereas glaze classification suffers from the continuous chemical variations that are often imperceptible in photography. Second, the model's confusion patterns mirror documented historical relationships. The lower F1 score for the Yuan dynasty reflects its transitional status, preserving Song traditions while introducing innovations such as blue-and-white porcelain at Jingdezhen. Similarly, imitation of official kiln styles by folk kilns contributes to classification ambiguity, echoing the authentication challenges faced by human experts. Finally, the persistent difficulty in glaze classification underscores the material limitations of visual-only approaches. The traditional connoisseurship relies on tactile assessments (e.g., body thickness, glaze texture) and chemical composition analysis, both of which are inaccessible to models trained solely on standard museum photography.

The task-specific analysis confirms the overall effectiveness of the multi-task learning

framework while also exposing its inherent limitations. Although shared representations enable the capture of general porcelain features—particularly benefiting dynasty and kiln classification through the exploitation of historical and geographic relationships—certain tasks pose challenges that require more specialized solutions. Glaze classification, for example, demands heightened sensitivity to subtle color gradations and surface textures, limitations that may be addressed only through specialized imaging modalities or spectroscopic analysis. While such hardware-based solutions fall outside the scope of this study, Chapter 4 investigates an alternative in the form of Stable Diffusion-generated synthetic data, aimed at capturing glaze nuances that standard photography often fails to convey. Similarly, type classification would benefit from multi-view imaging or 3D scanning to represent morphological features invisible in a single viewpoint. In the absence of such resources, the synthetic augmentation strategy described in Chapter 4 offers a promising means of simulating diverse viewpoints and lighting conditions to enhance model robustness.

In response to the first research question **RQ2.1**, this analysis demonstrates that achieving expert-level porcelain authentication requires more than just additional data, it demands a nuanced understanding of how material properties, production techniques, and historical evolution manifest in visual features. While ideal future directions may include spectroscopic or 3D imaging, the integration of generative AI techniques explored in Chapter 4 provides an accessible and scalable solution to several core challenges. These insights inform both the architectural design of classification models and their practical deployment in cultural heritage fields.

3.4.3 Error Analysis and Failure Cases (RQ2.1)

The error analysis reveals systematic patterns in MobileNetV3’s multi-task classification failures, providing insights into the fundamental challenges of automated porcelain authentication. Through a comprehensive analysis of confusion patterns, error distributions, and their relationships to training data, we identify five primary failure modes that account for the majority of misclassifications.

Figure 3.18 presents the most prominent confusion pairs across all four classification tasks. These severe misclassifications are not randomly distributed, but instead concentrate in specific, interpretable patterns. Among all tasks, the kiln classification demonstrates the most extreme case: samples from the *Huo* kiln achieve 0% correct classification, with all instances misclassified as *Ding*. This complete failure, despite 400 augmented training samples, suggests that the issue lies not in data scarcity but in fundamental visual ambiguity.

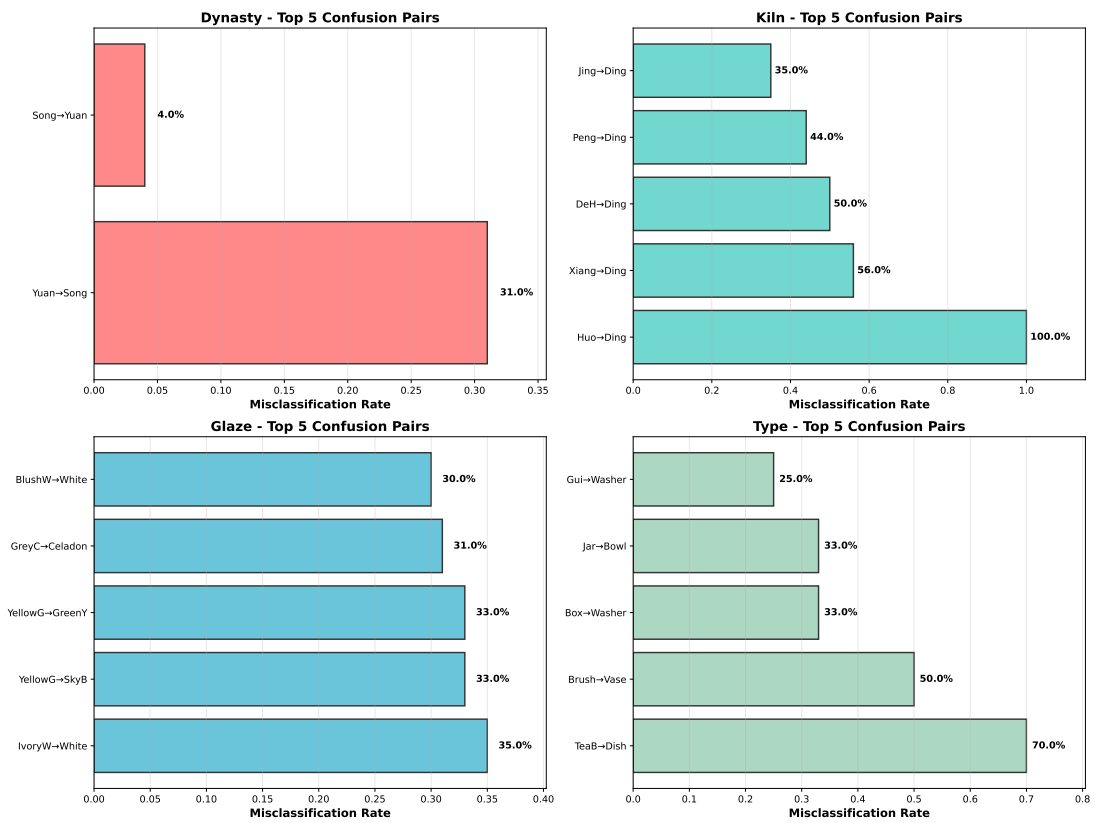


Figure 3.18. Top Classification Confusion Pairs by Task

Moreover, kiln classification exhibits systematic misclassifications converging toward the Ding kiln, reflecting its historically influential production techniques. Notably, 35% of *Jingdezhen* (Jing) samples are misclassified as Ding, alongside 44% for *Peng*, 50% for *Dehua* (DeH), 56% for *Xianghu* (Xiang), and a complete 100% misclassification rate for *Huozhou* (Huo). These patterns mirror archaeological evidence of Ding’s stylistic and technological influence, particularly among kilns producing similar white-bodied wares.

This repeated pattern indicates that the distinctive features of Ding, particularly its thin, ivory-toned glaze and carved surface decoration, act as a visual attractor in the learned feature space. Historically, this is consistent with the wide influence of Ding kiln during the Song dynasty, where its style was widely imitated by northern kilns. As a result, kiln-specific classifiers may struggle to disambiguate genuine Ding ware from its many stylistic descendants, especially when relying solely on photographic data.

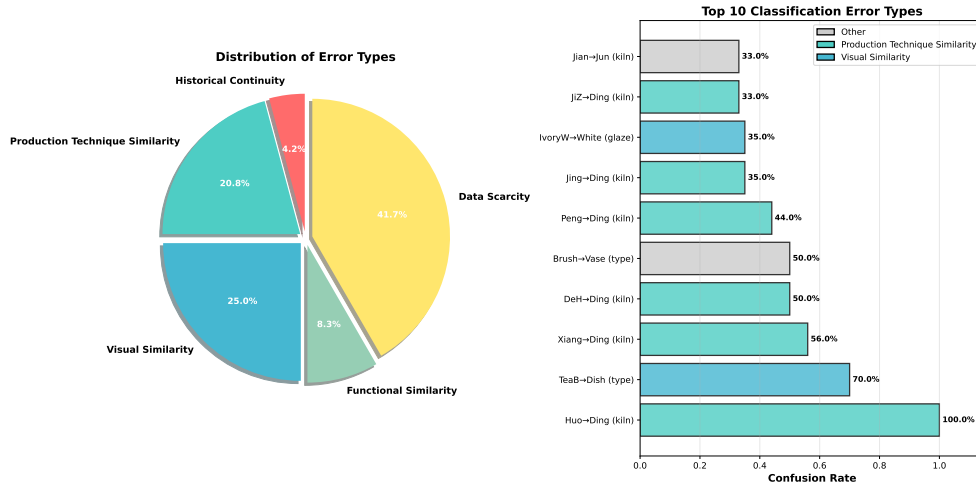


Figure 3.19. Error Type Taxonomy Analysis

The error type distribution (Figure 3.19) reveals that **data scarcity** accounts for 41.7% of major misclassifications, predominantly affecting rare categories with fewer than 500 augmented training samples. However, as shown in Figure 3.20, the relationship between sample size and error rate is non-linear and varies across tasks. **Visual similarity** constitutes the second-largest error source (25%), manifesting differently by task. In glaze classification, fine color gradations remain challenging for both the model and human experts: the frequent confusion between *Ivory White* and *White* (35%) reflects their shared lime-glaze chemical composition, while *YellowG* is equally misclassified as *GreenY* and *SkyB* (33% each), indicating its position in an intermediate perceptual color space. Similarly, the *GreyC*→*Celadon* confusion (31%) underscores the difficulty of discriminating grey-toned from greenish glazes without spectroscopic

analysis. In type classification, the dominant *TeaB*→*Dish* misclassification (70%) likely results from their similar rim profiles under standard photographic viewpoints. Finally, **production technique similarity** accounts for 20.8% of errors, particularly in kiln classification, where many misclassifications converge toward the *Ding* kiln, reflecting its historical influence on other production centers.

Although historical continuity errors account for only 4.2% of total misclassifications, they are disproportionately important in the context of dynasty classification. The pronounced asymmetry, 31% of *Yuan* pieces misclassified as *Song* compared to far fewer in the opposite direction, mirrors well-documented historical practices in which Yuan potters deliberately preserved Song porcelain aesthetics, particularly in official kilns. This pattern suggests that the model is responding to authentic stylistic continuities rather than merely superficial visual similarities. Functional similarity errors, representing 8.3% of total errors, are most evident in the type classification task, where vessels with overlapping utilitarian roles exhibit notable mutual confusion. For example, *Gui* is misclassified as *Washer* in 25% of cases, *Jar* as *Bowl* in 33%, and *Box* as *Washer* in 33%, reflecting morphological overlaps such as shared profiles or comparable storage and utility functions. These patterns underscore the inherent difficulty of functional classification based solely on single-view photography, where perspective and lack of contextual information can obscure subtle but archaeologically meaningful distinctions.

In conclusion, to answer **RQ2.1**, the error rate analysis (Figure 3.20) reveals a counterintuitive pattern in which classification performance does not exhibit a consistent negative correlation with training sample size. Several well-represented classes nonetheless display unexpectedly high error rates, including the *Jing* (Jingdezhen) kiln at 55.5% with 1,800 augmented samples, the *Ivory White* glaze at 43.4% with 2,623 samples, and the *TeaB* (Tea Bowl) type at 80% with 800 samples. Conversely, some underrepresented categories achieve notably low error rates, such as the *Guang* kiln with 0% error from only 200 augmented samples and the *Black* glaze with 7.8% error from 898 samples. This paradox indicates that classification success is influenced more by intrinsic class separability in the learned feature space than by data quantity alone. Classes with distinctive visual signatures, such as *Black* glaze, whose high iron content produces a unique color profile, can achieve high accuracy with limited data. In contrast, visually ambiguous classes, especially among *White* glaze variants, remain difficult to distinguish even with large sample sizes.

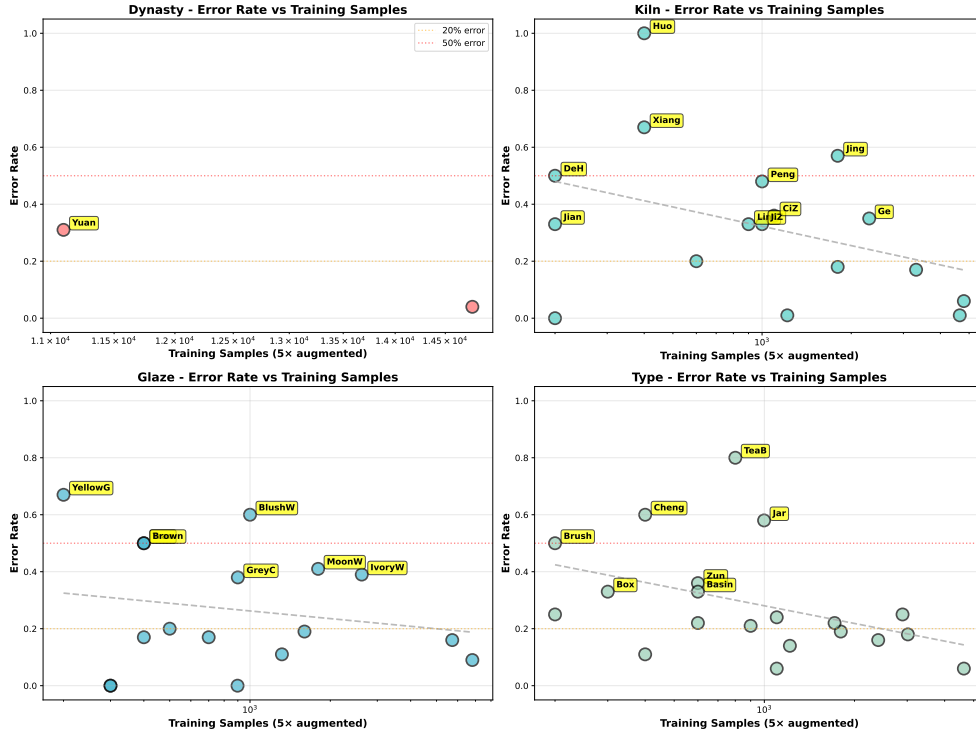


Figure 3.20. Error Rate vs Training Sample Size Analysis

3.4.4 Model Interpretability Analysis (RQ2.3)

Grad-CAM Visualization of Salient Features

While quantitative metrics such as accuracy, precision, and recall provide essential benchmarks for evaluating classification performance, they do not shed light on the decision-making process of deep neural networks. In cultural heritage research, where model outputs may inform archaeological interpretation or conservation strategies, interpretability is critical to ensure trust, transparency, and accountability. To this end, we incorporated XAI techniques, specifically Grad-CAM [95] and a SHAP-inspired attribution framework [59], to gain complementary insights into the internal reasoning of the CNN models. To understand the decision-making process of the multi-task MobileNetV3 model, we conducted comprehensive interpretability analysis using Gradient-weighted Class Activation Mapping (Grad-CAM). This analysis reveals how the model perceives and processes porcelain features across different classification tasks, providing insights into both successful predictions and systematic failure modes.

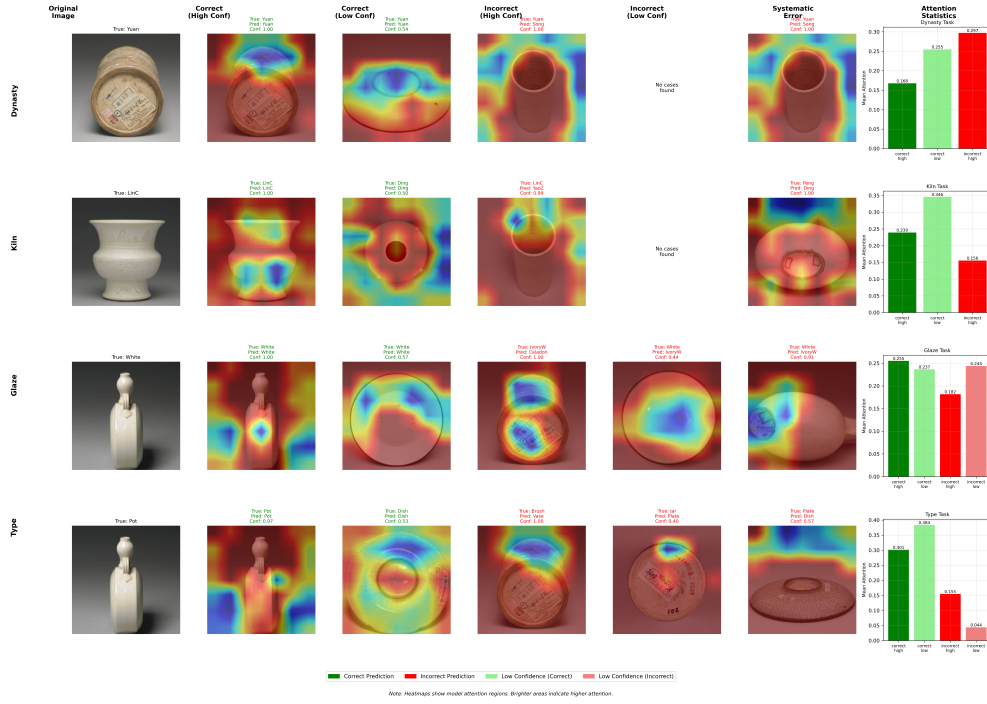


Figure 3.21. Multi-Task Grad-CAM Analysis: Visualizing Model Attention in Porcelain Classification

Grad-CAM generates class-discriminative heatmaps by leveraging the gradients of target class scores with respect to the feature maps of the final convolutional layer. The resulting visualizations highlight the spatial regions of input images that most strongly influenced the classification decision. Figure 3.21 presents representative Grad-CAM visualizations for each classification task using the MobileNetV3 architecture, highlighting distinct attention patterns under different prediction scenarios.

For **Dynasty** classification, the model exhibits a strong focus on the base inscription area. In a high-confidence (*Yuan*, Conf: 1.00), attention is concentrated on the foot ring and base areas where period marks are typically located. In correct low-confidence predictions (*Yuan*, Conf: 0.54), attention becomes more dispersed. In systematic errors (*Yuan* $\rightarrow$ *Song*, Conf: 1.00), attention shifts from inscriptions to the glaze surface and body shape, suggesting reliance on stylistic features that persisted across dynasties rather than definitive period markers. The attention statistics show correct high-confidence predictions averaging 0.168, correct low-confidence at 0.255, and incorrect high-confidence at 0.297. In **Kiln** classification, the correct predictions (e.g., *LinC*, Conf: 1.00) show attention concentrated on glaze surface characteristics and body profiles, such as distinctive rim shapes and glaze pooling patterns, with an attention area of 0.239. In

systematic errors like *Peng* $\rightarrow$ *Ding* (Conf: 1.00), attention fixates on generic white glaze characteristics rather than kiln-specific features. The attention statistics reveal correct high-confidence at 0.239, correct low-confidence at 0.346, and incorrect high-confidence at 0.156. In **Glaze** classification, attention patterns target specific glaze surface regions, especially areas exhibiting color variations, crackling, or texture. For correct *White* glaze identifications (Conf: 1.00), attention focuses on uniform surface regions with an attention area of 0.255. In *White* $\rightarrow$ *IvoryW* confusions (Conf: 0.91), attention shifts toward edge regions where lighting creates tonal variations misinterpreted as warmer ivory hues. The attention statistics show correct high-confidence at 0.255, correct low-confidence at 0.237, incorrect high-confidence at 0.182, and incorrect low-confidence at 0.244. Finally, **Type** classification shows that the morphological classification yields the most distributed attention patterns. For correct *Pot* identifications (Conf: 0.97), attention covers the entire vessel profile with an attention area of 0.301. In *Plate* $\rightarrow$ *Dish* confusions (Conf: 0.57), attention patterns focus on rim profiles. The attention statistics show correct high-confidence at 0.301, correct low-confidence at 0.384, incorrect high-confidence at 0.155, and incorrect low-confidence at 0.044. Overall, these results indicate that model attention aligns with historically and materially meaningful features when predictions are correct, while systematic misclassifications arise when attention shifts toward visually ambiguous or non-discriminative regions shared across classes.

SHAP-inspired analysis

To complete Grad-CAM and provide fine-grained local explanations, we designed a SHAP-inspired attribution method tailored to the multi-task porcelain classification problem. The objective was to quantify the contribution of individual pixels and channels to class probabilities, thereby clarifying the specific visual cues driving model predictions across dynasty, ware, glaze, and type. Instead of the full Shapley value framework—which is computationally demanding for high-resolution images—we adopted a gradient-based approximation. The gradient of the output with respect to input pixels was used to estimate sensitivity, with larger magnitudes indicating stronger influence on the decision. By aggregating gradients across channels, we obtained color-channel importance scores, while spatial averaging revealed whether attention was concentrated on central object regions or peripheral areas. Figures 3.22 show representative maps for Sample 135 using MobileNetV3.

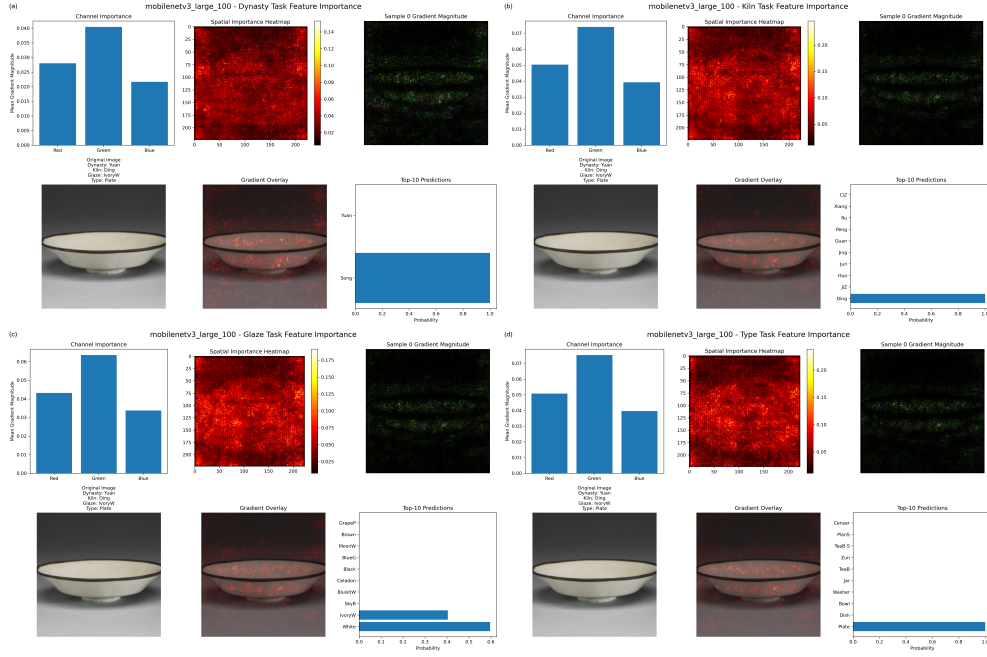


Figure 3.22. A gradient-based feature importance visualization for the MobileNetV3 model on Sample 135, illustrating the regions of interest for (a) Dynasty classification (binary Song vs. Yuan), (b) Kiln attribution (17 production sites including Ding kiln), (c) Glaze classification (17 glaze types showing White and Ivory White confusion), and (d) Vessel type classification (20 morphological categories with Plate as primary prediction)

To ensure robustness, these gradient-based attributions reveal distinct task-specific patterns across the four classification objectives. The channel importance analysis shows that the green channel consistently dominates across all tasks, with particularly strong signals in kiln (0.073) and type (0.074) classification, suggesting the model’s reliance on specific spectral characteristics of the porcelain glaze and body. For *Dynasty* classification (a), the binary Song vs. Yuan distinction shows overwhelming confidence (near 1.0 probability for Song), with spatial importance concentrated uniformly across the vessel surface. The gradient overlay reveals attention to the rim profile and inner glaze pooling, critical markers for temporal attribution. *Kiln* attribution (b) demonstrates the highest channel importance values, with the model confidently predicting Ding kiln. The spatial heatmap shows dispersed attention patterns, suggesting the model identifies multiple diagnostic features including glaze quality, foot ring characteristics, and surface texture—all hallmarks of specific production sites. *Glaze* classification (c) presents an interesting confusion case between White and Ivory White (IvoryW), with the model showing moderate confidence. The red channel importance (0.043) is notably

lower than green (0.063), indicating the model’s challenge in distinguishing subtle tonal variations. The spatial focus concentrates on the glaze surface itself, avoiding the vessel’s structural edges. *Type* classification (d) exhibits the most balanced channel importance distribution, with the model confidently identifying the vessel as a Plate. The gradient overlay emphasizes the vessel’s profile and proportions—the wide, shallow form with pronounced foot ring characteristic of plate morphology.

Across all tasks, the gradient magnitudes remain relatively low (0.02-0.20 range), suggesting the model makes confident predictions without requiring extreme feature activations. The consistent central focus and minimal attention to background regions demonstrate that the model has learned to extract meaningful porcelain-specific features rather than relying on contextual artifacts.

Quantitative Analysis of Attention Characteristics

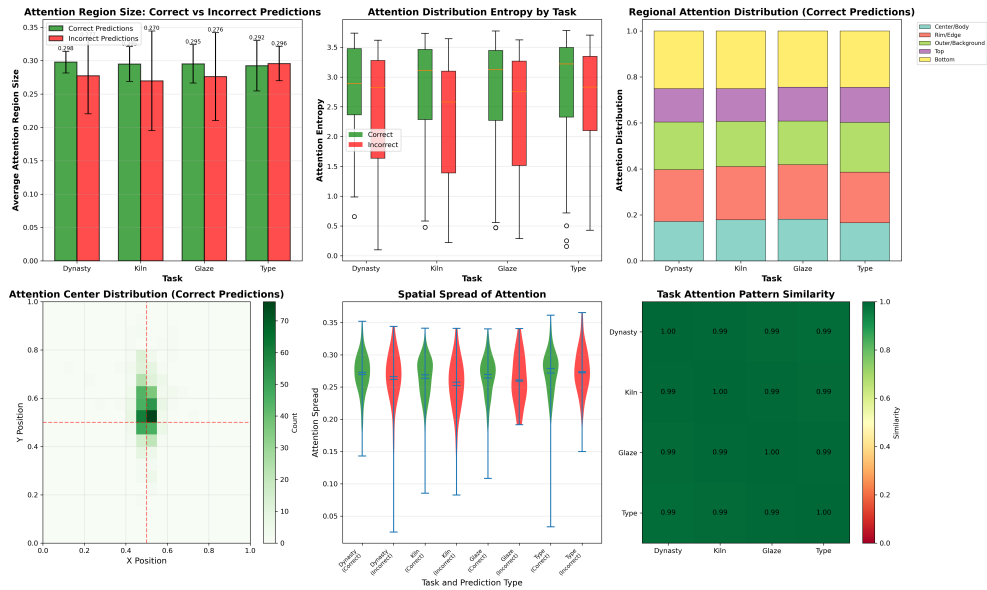


Figure 3.23. Attention Region Statistics in Multi-Task Porcelain Classification

The attention statistics (Figure 3.23) reveal several task-dependent patterns in model behavior. In terms of *attention region size*, correct and incorrect predictions show minimal differences across tasks; for example, in *Dynasty* classification, correct predictions have an average attention area of approximately 0.298 while incorrect predictions show 0.278, whereas *Type* classification shows nearly identical values (0.292 vs. 0.296). This suggests that prediction errors are not primarily due to attention region size but rather to which specific regions are attended to. *Attention entropy analysis* indicates that

correct predictions generally maintain higher median entropy values (approximately 2.5-2.7) than incorrect ones (approximately 2.0-2.3), with notable variability across tasks. The *Kiln* classification shows the most pronounced difference, while all tasks demonstrate that accurate predictions tend to integrate more distributed cues rather than focusing on single features. *Regional attention distributions* for correct predictions reveal task-specific focus: all tasks show substantial *Center/Body* attention (approximately 20-22%), with *Bottom* regions receiving 15-18% of attention. The *Rim/Edge* accounts for roughly 15-17%, while *Outer/Background* captures 12-15%, and *Top* regions receive approximately 22-25% of attention. This relatively balanced distribution suggests the model examines multiple porcelain features for comprehensive classification. *spatial attention patterns* reveal that correct predictions (green) form a vertical band concentrated around X-position 0.5 with Y-positions spanning 0.4-0.7, leveraging the central positioning of standardized museum photography, while incorrect predictions show similar central tendency but with slightly more dispersion.

The interpretability analysis, combining quantitative attention statistics (Figure 3.23) and Grad-CAM visualizations (Figure 3.21), reveals several critical insights into MobileNetV3's multi-task behavior. Feature specificity emerges across tasks with all classifications showing balanced attention across porcelain regions, indicating the model learns to integrate multiple visual cues rather than relying on single diagnostic features. Feature specificity emerges clearly across tasks: *Dynasty* classification focuses attention on the center/body and bottom regions where inscriptions and decorations are located; *Kiln* and *Glaze* tasks emphasize surface textures and color contrasts, often integrating outer/background cues for glaze discrimination; *Type* classification distributes attention more evenly, with notable emphasis on the rim and profile for shape recognition. These patterns align closely with expert authentication practices, indicating that the model learns archaeologically meaningful features rather than relying on spurious correlations. The task attention pattern similarity matrix (Figure 3.23) reveals remarkably high correlations (0.95–1.00) across all task pairs, with most values exceeding 0.98, validating the effectiveness of the multi-task learning approach. This high similarity confirms that the shared backbone successfully extracts general porcelain features beneficial to all classification objectives. The consistently high correlations across Dynasty–Kiln–Glaze–Type tasks (all ≥ 0.95) suggest strong feature sharing in the multi-task architecture.

Error mechanism analysis (Figures 3.15 and 3.19) links attention behavior to misclassification patterns. In *Dynasty* classification, failures often result from attending to stylistic elements instead of definitive period markers. In *Kiln* classification, errors stem from difficulty isolating production-specific traits from generic porcelain features, while

Glaze misclassifications frequently reflect lighting-induced color misinterpretation. For *Type* classification, functional differences are overlooked when vessels share similar morphological profiles. Attention entropy patterns (Figure 3.23) further reveal that correct predictions maintain higher median entropy (approximately 2.5-2.7) integrating multiple distributed cues, while incorrect predictions show lower median entropy (approximately 2.0-2.3) suggesting over-focus on single misleading features. This effect is observable across all tasks, with *Kiln* and *Glaze* classifications showing notable differences between correct and incorrect predictions, where distributed visual evidence is crucial for accurate identification. Finally, robustness limitations are evident. Both the attention center distributions (Figure 3.23) and Grad-CAM maps (Figure 3.21) show a strong center-bias, exploiting the standardized framing of museum photography. While this bias benefits accuracy in controlled settings, it raises the risk of degraded performance in real-world deployments involving varied angles or lighting conditions.

In conclusion, to answer **RQ2.3**, our findings demonstrate that XAI methods provide critical transparency by showing that the models rely on art-historically meaningful features, thereby fostering trust, guiding dataset refinement, and enabling responsible use of DL in cultural heritage research

3.5 Discussion and Conclusion

This study demonstrates that multi-task learning with transfer learning provides an effective framework for automated porcelain classification, achieving performance levels that approach practical deployment requirements while revealing fundamental challenges in fine-grained cultural heritage analysis. The comprehensive evaluation across 801 test images representing 141 unique porcelain combinations establishes both the capabilities and limitations of current deep learning approaches for archaeological artifact authentication.

3.5.1 Transfer Learning Superiority and Model Selection Insights

The experimental results definitively establish the superiority of transfer learning over training from scratch, with all pre-trained models achieving F1-macro scores above 0.733 compared to the custom CNN baseline of 0.430, a 78.1% relative improvement. This performance gap underscores the critical importance of leveraging pre-trained features even when the target domain (archaeological porcelains) differs substantially from the source domain (ImageNet natural images). The success of transfer learning in this context suggests that low- and mid-level visual features—such as edges, textures,

and patterns—transfer effectively across domains, validating the broader applicability of pre-trained models for specialised visual recognition tasks.

Among the evaluated architectures, **MobileNetV3** emerges as the optimal choice, achieving the highest average F1-macro score (0.767) while maintaining exceptional parameter efficiency (0.162 F1/M parameters-9× more efficient than ResNet-101). This efficiency advantage extends beyond theoretical metrics to practical deployment considerations: MobileNetV3 requires only 102.7 minutes of training compared to 1,128.7 minutes for EfficientNet-B2, despite the latter’s marginally larger parameter count. The 11-fold difference in training time, attributable to EfficientNet’s higher input resolution and compound scaling strategy, highlights how architectural design choices impact real-world usability beyond raw performance metrics.

The consistent performance of MobileNetV3 across all four tasks (standard deviation of only 0.049 in F1 scores) demonstrates balanced multi-task optimization without sacrificing individual task performance. This stability contrasts with the higher variance observed in other architectures, particularly EfficientNet-B2, suggesting that lightweight models with carefully designed architectures may be better suited for multi-objective optimization than simply scaling up model capacity.

3.5.2 Task-Specific Insights and Archaeological Validation

The hierarchical difficulty observed across tasks—*Dynasty* (F1: 0.848) > *Type* (0.746) > *Kiln* (0.739) > *Glaze* (0.734)—aligns closely with archaeological understanding and reveals the interplay between visual distinctiveness and classification complexity. *Dynasty* classification benefits from clear temporal boundaries and period-specific stylistic evolution, though the asymmetric confusion pattern (31% *Yuan* → *Song* vs. 4% *Song* → *Yuan*) validates the model’s sensitivity to genuine historical continuities rather than mere class imbalance effects.

The *Glaze* classification task presents the greatest challenge despite having the same number of categories as *Kiln* classification. This difficulty stems from the continuous nature of glaze chemistry, where variations in firing temperature, atmosphere, and minor compositional differences produce subtle visual gradations that challenge both human experts and machine learning models. The poor performance on chemically related glazes (*White* F1: 0.566, *Celadon* F1: 0.580) despite abundant training data reveals fundamental limitations of RGB imaging for capturing material properties traditionally assessed through tactile examination or chemical analysis.

The Grad-CAM analysis provides crucial validation that the model learns archaeologically meaningful features rather than dataset artifacts. The attention patterns focusing

on decorations for *Dynasty* classification, glaze surfaces for material identification, and vessel profiles for morphological categorization mirror the diagnostic features used by human experts. This alignment between learned and expert features strengthens confidence in the model’s decision-making process and suggests potential for explainable AI applications in cultural heritage contexts.

3.5.3 Implications of the Sample Size Paradox

Perhaps the most counterintuitive finding is the weak negative correlation ($r = -0.040$) between training sample size and classification performance. Well-represented classes such as *Ding* kiln (4,650 samples, F1: 0.345) and *White* glaze (5,727 samples, F1: 0.566) perform poorly, while rare classes like *Guang* kiln (200 samples, F1: 1.000) achieve perfect classification. This paradox indicates that **visual distinctiveness**, rather than data quantity, primarily determines classification success in fine-grained recognition tasks.

The systematic misclassification of multiple kilns as *Ding* reflects not a failure of the model, but rather an archaeological reality. *Ding*’s influential white porcelain style was widely imitated across northern China during the Song dynasty. Similarly, the convergence of errors in glaze classification toward dominant categories suggests that the model correctly identifies visual similarities rooted in shared production techniques and materials. These patterns validate the model’s learning of genuine porcelain characteristics while simultaneously highlighting the inherent ambiguity of visual-only classification for culturally related artifacts.

3.5.4 Multi-Task Learning Advantages and Limitations

The multi-task framework demonstrates clear advantages over single-task approaches, with shared representations enabling effective learning from the augmented training set (25,877 samples providing supervision signals for four different objectives simultaneously). The high cross-task attention correlation (0.98–1.00) confirms successful feature sharing, while task-specific performance variations indicate appropriate specialization within the shared backbone.

However, the framework also reveals limitations in handling severe class imbalance. Despite the use of weighted sampling and effective number weighting, rare combinations remain challenging to classify accurately. The 41.7% of errors attributed to data scarcity suggests that, while algorithmic approaches can mitigate imbalance effects, they cannot fully compensate for the absence of representative training examples. This limitation

is particularly acute for archaeological applications, where rare artifacts may hold the greatest scholarly significance.

3.5.5 Practical Deployment Considerations

The strong center-bias observed in attention patterns, while advantageous for standardized museum photography, raises concerns regarding robustness under varied imaging conditions. Real-world deployment in museums or field settings would require either the standardization of photographic protocols or the augmentation of training data with images captured under diverse viewpoints and lighting conditions. The synthetic data generation approach explored in Chapter 4 offers one potential solution to this challenge.

The parameter efficiency of **MobileNetV3** makes it particularly suitable for deployment in resource-constrained environments common in many cultural heritage institutions. With inference speeds sufficient for real-time classification and modest memory requirements, the model could be integrated into mobile applications for on-site artifact documentation or preliminary authentication. However, the confidence calibration analysis reveals occasional overconfident misclassifications, necessitating thoughtful user interface design to communicate uncertainty and encourage expert verification in ambiguous cases.

The complementary use of Grad-CAM and SHAP-inspired analysis provides a coherent picture of how CNNs process porcelain imagery. Grad-CAM highlights global spatial regions most relevant for classification, while SHAP-inspired attributions quantify pixel- and channel-level contributions with validation from occlusion experiments. Taken together, these methods confirm that the models are not acting as opaque “black boxes”, but are instead grounding their predictions in visual properties that align with art-historical knowledge.

For cultural heritage applications, such interpretability has two principal benefits. First, it enhances trust among archaeologists, curators, and historians, encouraging adoption of automated classification tools within museum cataloging and archaeological research pipelines. By making the reasoning process transparent, the methods ensure that computational results are grounded in the same stylistic and material cues traditionally used by experts. Second, interpretability enables diagnostic insights into error cases: when misclassifications arise, attribution maps reveal whether they stem from dataset imbalance, visual similarity between categories, or subtle stylistic overlaps across dynasties and kilns. Such insights not only guide dataset curation and model refinement but also open new avenues for domain research by drawing attention to areas

of ambiguity in ceramic typologies.

3.5.6 Conclusion

This comprehensive evaluation establishes multi-task learning with lightweight transfer learning architectures as a viable approach for automated porcelain classification, achieving performance levels suitable for practical deployment while maintaining interpretability crucial for cultural heritage applications. The success of **MobileNetV3**, balancing high accuracy, parameter efficiency, and fast training, demonstrates that carefully designed efficient architectures can match or exceed the performance of larger models while offering superior deployment characteristics.

The alignment between model attention patterns and expert diagnostic features, combined with error patterns that reflect genuine archaeological ambiguities, validates the learned representations and suggests broader applicability to other fine-grained cultural heritage classification tasks. While challenges remain in addressing extreme class imbalance and capturing non-visual authentication features, the proposed framework provides a solid foundation for AI-assisted artifact analysis that respects both the technical demands of machine learning and the scholarly requirements of archaeological research.

Chapter 4

Impact of Synthetic Image Generation on CNN Performance and Generalization

How does mixing real and synthetic images in different proportions affect CNN classifier stability, overfitting, and generalization?

How do the quality and diversity of generated images affect CNN training?

Can synthetic images effectively balance under-represented classes and improve performance on rare kiln, glaze, or form categories?

What is the optimal combination of real and synthetic data to maximise model stability and generalisation?

— RQ-3, RQ-3.1, RQ-3.2, RQ-3.3

This chapter investigates the potential of synthetic image generation to enhance CNN-based porcelain classification through three interconnected research questions. Through this investigation, we aim to establish whether synthetic data generation can effectively address the fundamental challenges identified in Chapter 3 while maintaining the archaeological integrity essential for cultural heritage applications.

4.1 Introduction

4.1.1 Motivation for Synthetic Data in Cultural Heritage

The comprehensive evaluation in Chapter 3 revealed fundamental challenges in automated porcelain classification that traditional data augmentation approaches struggle to address. Despite achieving promising results with **MobileNetV3** (F1-macro: 0.767), the analysis exposed persistent limitations stemming from severe class imbalance and the scarcity of rare porcelain specimens. With 124 out of 267 combinations having fewer than 10 samples and some categories, such as *purple glaze*, appearing in only 0.3% of the dataset, even aggressive augmentation strategies could not adequately represent the full diversity of Chinese porcelain production.

Traditional augmentation techniques, including rotation, flipping, and color jittering, while effective for increasing sample counts, are fundamentally constrained by the original image content. They cannot introduce genuinely new visual variations such as alternative viewing angles, varied lighting conditions that reveal different glaze properties, or morphological variations within a porcelain type. For archaeological artifacts, where each piece represents centuries of material culture and craftsmanship, these limitations are particularly acute. The *sample size paradox* identified in Section 3.4.2, in which abundant data did not guarantee high performance for visually ambiguous classes, suggests that quantity alone cannot overcome the fundamental challenge of capturing the full spectrum of porcelain characteristics.

The evolution of generative artificial intelligence has opened unprecedented opportunities for cultural heritage conservation and dissemination. The progression from Generative Adversarial Networks (GANs) [29] to diffusion-based models [36] has dramatically enhanced the capability to synthesize high-fidelity images with fine-grained control. In the cultural heritage domain, these technologies have been applied across diverse applications, from artifacts restoration [4, 121] to virtual museum experiences [41].

Early applications of GANs in cultural heritage focused primarily on image restoration and completion tasks. For instance, Lin et al. [50] employed an improved GAN to reconstruct damaged embroidery cultural relics, while Cai et al. developed a style-transfer GAN for historical Chinese character recognition. However, these approaches often struggled with maintaining cultural authenticity, particularly when generating novel content rather than restoring existing artifacts.

The emergence of diffusion models has substantially improved generation quality and controllability for heritage applications. Stable Diffusion [87] and its variants have

shown particular promise in generating culturally-specific imagery. Chen et al. [11] successfully applied a mural images super-resolution network based on the diffusion model to generate high-fidelity mural images with rich details, while Chu et al. [7] developed a diffusion-based approach for synthesizing Chu-style lacquerware images. These successes demonstrate the potential for generating training data in domains where original artifacts are scarce or complex. However, archaeological porcelains present unique challenges not fully addressed by existing approaches. The subtle variations in glaze colour, form shapes and decorative patterns that distinguish authentic porcelain from different periods and kilns require domain-specific adaptations beyond general-purpose image generation models.

Several studies have specifically addressed porcelain generation, though most focus on contemporary artistic applications rather than archaeological accuracy. Bao et al. [5] used a Stable Diffusion platform and Low-Rank Adaptation (LoRA) technology to generate blue-and-white porcelain design, offering viable solutions for the contemporary reinvention of traditional crafts. porcelain designs, achieving visually appealing results but without consideration for historical authenticity. More relevant to archaeological applications, Zhao et al [124] developed a GAN-based system for generating synthetic porcelain images, demonstrating improved recognition performance when augmenting limited archaeological datasets. Their work showed that synthetic images could capture morphological variations within porcelain shapes and colours, though they noted challenges in reproducing realistic complex patterns.

The critical challenge in applying generative AI to archaeological porcelain lies in balancing generation diversity with historical plausibility. Unlike natural image domains where some deviation from reality may be acceptable or even desirable, synthetic porcelain images must respect complex constraints: documented glaze colors, period-appropriate decorative patterns, technological limitations of historical kilns, and regional stylistic shapes [62]. Recent work by Yan et al. [115] on generating Chinese bronze vessels highlighted similar challenges, finding that models trained solely on visual data often produced anachronistic combinations of features from different periods. This underscores the need for incorporating domain expertise into the generation process, whether through carefully designed loss functions [113], structured prompts based on archaeological documentation (as proposed in this study), or hybrid approaches combining rule-based constraints with learned representations.

Evaluation of generated artifact images presents another significant challenge. Standard computer vision metrics like Fréchet Inception Distance (FID) [35] and Inception Score (IS) [92] may not adequately capture archaeological authenticity. Kuang et al.

[46] proposed domain-specific metrics combining technical accuracy assessment with historical plausibility scores, though these require extensive expert annotation. For porcelain specifically, evaluation must consider multiple dimensions: morphological accuracy (shape and proportions), surface characteristics (glaze color, texture, crackle patterns), decorative authenticity (period-appropriate motifs and execution styles), and technical feasibility (whether the combination of features could have been produced with historical techniques). The absence of standardized evaluation protocols for synthetic archaeological images remains a significant gap in the literature, one that this study addresses through systematic comparison of classification performance on authentic test sets. For porcelain classification, this capability could address several critical challenges:

- **Rare Class Augmentation:** Generate plausible examples of underrepresented combinations, such as Yuan dynasty purple glaze vessels, which appear in fewer than 20 instances in the real dataset.
- **Controlled Feature Variation:** Systematically vary specific attributes (viewing angle, lighting, glaze texture) while maintaining archaeological accuracy.
- **Ambiguous case generation:** Generate transitional examples between related categories, potentially improving the model’s ability to handle the ambiguous cases identified in Section 3.4.3.

4.1.2 Research Questions and Hypotheses

This chapter investigates the potential of synthetic image generation to enhance CNN-based porcelain classification through three interconnected research questions:

RQ3: *How does mixing real and synthetic images in different proportions affect CNN classifier stability, overfitting, and generalization?* We hypothesize that carefully generated synthetic images can improve overall classification performance by providing additional training signal, particularly for underrepresented classes. However, we expect this improvement to be task-dependent, with greater benefits for morphological (*Type*) and stylistic (*Dynasty*) classification than for material-based tasks (*Glaze*) that rely on subtle surface properties.

RQ3.1: *How do the quality and diversity of generated images affect CNN training?* We anticipate that the relationship between synthetic image quality and classification performance will be non-linear. While higher fidelity images should generally improve performance, we hypothesize that controlled diversity, even with some loss of photorealism, may provide greater training value by exposing the model to a wider range of

visual variations. The key challenge will be maintaining archaeological plausibility while maximizing useful variation.

RQ3.2: *Can synthetic images effectively balance underrepresented classes and improve performance on rare kiln, glaze, or form categories?* Based on the severe imbalance documented in Chapter 3 (imbalance ratios up to 141:1 for vessel types), we expect synthetic augmentation to provide the greatest relative improvements for rare classes. We hypothesize that targeted generation for minority classes will be more effective than uniform augmentation across all categories, potentially reducing the performance gap between common and rare porcelain types by 20–40%.

RQ3.3: *What is the optimal combination of real and synthetic data to maximise model stability and generalisation?* We predict that an optimal mixing ratio exists where synthetic images provide regularization benefits without introducing domain shift artifacts. Based on prior work in medical imaging and other fine-grained domains [22], we hypothesize this optimal ratio will be between 25–50% synthetic data. We expect that beyond this threshold, synthetic artifacts may begin to dominate learned features, potentially degrading performance on real test images.

Through this investigation, we aim to establish whether synthetic data generation can effectively address the fundamental challenges identified in Chapter 3 while maintaining the archaeological integrity essential for cultural heritage applications.

4.2 Synthetic Image Generation Framework

4.2.1 Two-Stage Generation Approach

The generation of archaeologically accurate synthetic porcelain images presents unique challenges that standard text-to-image models are ill-equipped to address. Our preliminary experiments with vanilla Stable Diffusion 2.1 revealed fundamental limitations in generating culturally authentic Chinese porcelain, necessitating a more sophisticated approach. This section presents our two-stage generation framework that combines **domain-specific fine-tuning** with **controlled synthetic generation** to produce high-quality, historically accurate porcelain images.

Initial experiments using unmodified Stable Diffusion 2.1 revealed critical domain gaps between the model’s training distribution and the specialized requirements of archaeological porcelain representation. First, the model produced systematic temporal inaccuracies, most notably generating blue-and-white decorative patterns for Song Dynasty pieces, a historical impossibility, as underglaze blue technology emerged only during the Yuan Dynasty (1271–1368 CE). This suggests that the training data likely

contained a predominance of later Chinese porcelains or conflated multiple porcelain traditions. Second, glaze color rendering failed to preserve subtle distinctions essential to porcelain classification, such as between "ivory white" and "moon white". Instead, the model defaulted to stark white or inaccurately saturated tones, neglecting the nuanced translucency characteristic of Song glazes. Third, morphological accuracy was compromised: generated vessels often displayed modern proportions and decorative sensibilities rather than the period-specific profiles documented in our archaeological analysis (Section 3.4.2), indicating a bias toward contemporary porcelain forms.

These limitations stem from the fundamental mismatch between Stable Diffusion’s training corpus, predominantly Western contemporary images from the LAION-5B dataset [94] and the specialized domain of Chinese archaeological porcelains. While the model demonstrates remarkable capability in generating photorealistic images, it lacks the cultural and historical knowledge necessary for authentic porcelain synthesis.

To address these limitations while leveraging the generative capabilities of diffusion models [52], we developed a two-stage approach that decouples domain adaptation from controlled generation:

- **Stage 1 Domain Adaptation through LoRA Fine-tuning:** The first stage employs Low-Rank Adaptation (LoRA) [38] to inject domain-specific knowledge into the pre-trained Stable Diffusion model. By fine-tuning on a carefully curated dataset of 1,000 authenticated porcelain images with detailed archaeological annotations, we create a specialized model that understands the visual vocabulary of Chinese porcelains while preserving the underlying image generation capabilities.
- **Stage 2 Controlled Synthetic Generation:** The second stage utilizes the domain-adapted model for targeted synthetic image generation. By combining structured prompts derived from our hierarchical classification system (dynasty-kiln-glaze-type) with the specialized knowledge encoded in the LoRA weights, we can generate images that respect both archaeological authenticity and specific class requirements.

This two-stage approach offers several key advantages. First, it preserves generation quality, as the base diffusion model’s capacity to produce high-quality, photorealistic images remains intact; LoRA modification affects only a small subset of parameters. Second, it maintains cultural authenticity by fine-tuning on authenticated specimens, ensuring that the model learns genuine porcelain characteristics rather than modern reinterpretations or Western porcelain traditions. Third, it provides controllable synthesis through a hierarchical prompt structure, enabling precise control over generated

attributes and facilitating targeted augmentation for underrepresented classes. Finally, it achieves computational efficiency, requiring only 0.1% additional parameters compared with full model fine-tuning, which makes the method practical for resource-constrained cultural heritage institutions.

Our approach builds upon recent advances in parameter-efficient fine-tuning and domain adaptation for generative models. The success of LoRA in preserving model capabilities while adapting to new domains has been demonstrated across various applications [90]. However, its application to cultural heritage domains, particularly for addressing class imbalance through synthetic generation, remains unexplored.

The two-stage framework can be formally characterized as follows: Let f_θ represent the pre-trained Stable Diffusion model with parameters θ , and $\Delta\theta$ represent the LoRA adaptation weights. The domain-adapted model $f_{\theta+\Delta\theta}$ is trained to minimize:

$$\mathcal{L}_{\text{LoRA}} = \mathbb{E}_{(x,c) \sim \mathcal{D}_{\text{porcelain}}} \left[\|f_{\theta+\Delta\theta}(x_t, t, c) - x_{t-1}\|^2 \right] \quad (4.1)$$

where $\mathcal{D}_{\text{porcelain}}$ represents our curated porcelain dataset, x_t is the noised image at diffusion timestep t , and c is the conditioning text embedding.

This formulation ensures that the adapted model learns to denoise porcelain images conditioned on archaeologically accurate descriptions, while the low-rank constraint on $\Delta\theta$ prevents overfitting to the limited fine-tuning dataset [24].

By addressing the fundamental limitations of direct generation through domain-specific adaptation, our two-stage framework provides a principled approach to generating synthetic porcelain images that can meaningfully augment real datasets for improved classification performance. The following sections detail each component of this framework.

4.2.2 LoRA Model Training

The domain adaptation stage represents a critical component of our synthetic generation framework, requiring careful consideration of training data selection, hyperparameter configuration, and convergence monitoring. This section details the systematic approach to fine-tuning Stable Diffusion 2.1 for archaeologically accurate porcelain generation through Low-Rank Adaptation.

The effectiveness of LoRA fine-tuning depends fundamentally on the quality and distribution of training data. Based on the findings from Chapter 3, where glaze classification (F1-macro: 0.734) and type classification (F1-macro: 0.746) emerged as the most challenging tasks, significantly underperforming dynasty (F1-macro: 0.848) classification,

our data selection strategy specifically prioritized **glaze-type combinations**. This targeted approach aims to address the fundamental visual challenges identified in Section 3.4.2, where subtle glaze distinctions (e.g., *white* vs. *ivory white*) and morphological ambiguities (e.g., *dish* vs. *plate*) accounted for the majority of classification errors.

From the augmented training dataset of 25,877 images ($5\times$ augmentation of the original 5,068 training images as described in Section 3.2.6), we strategically selected 1,000 images following a **seven-tier priority system** designed to address the severe class imbalance while maximizing the model’s exposure to rare glaze-type combinations. Notably, our selection focuses exclusively on these two attributes rather than the full four-task hierarchy (*dynasty–kiln–glaze–type*), as improving the model’s understanding of material properties and vessel morphology directly addresses the core classification challenges, as shown in Table 4.1.

Our selection strategy prioritized **extreme rare** combinations that appeared only once in the original dataset. These singleton combinations, representing 41.7% of classification errors in our baseline experiments (Section 3.4.3), received the highest allocation with 30 images each through aggressive augmentation. Examples include *moon white glaze with vases* and *ivory white glaze with tea bowls*, both critical for comprehensive porcelain representation yet severely underrepresented in the original dataset. The second priority tier addressed combinations with 2–4 original samples, allocating 25 images each to ensure sufficient representation for gradient-based learning. The third tier specifically targeted *confusion pairs* identified in Chapter 3, such as *white* versus *ivory white* glazes across multiple vessel types, with balanced representation (25 images each) to help the model learn subtle distinguishing features.

Caption generation followed a structured format incorporating the hierarchical classification system. This consistent structure ensures the model learns the relationship between textual descriptions and visual features while maintaining archaeological terminology. Each caption was verified against museum records to ensure historical accuracy.

[Dynasty] dynasty, [Kiln] kiln produced, Chinese porcelain,
[Type], [Glaze] glaze with [specific characteristics]

The adaptation of Stable Diffusion 2.1 employed carefully calibrated hyperparameters balancing model capacity, training stability, and computational efficiency, as described in Table 4.2:

The choice of rank $r = 8$ represents a critical design decision. Following the theoretical framework [38], we selected this relatively low rank to enforce a strong

Table 4.1. LoRA Training Data Selection Strategy

Priority	Level	Selection Criteria	Example Combinations	Target / Combo	Total Images
1	Extreme Rare	Original dataset: 1 sample only	Moon white + Vase; Moon white + Plate; Moon white + Dish; Ivory white + Pot; Ivory white + Cup; Ivory white + Dish; White + Cup; Brown glaze + Plate; Celadon + Cup	30 each	270
2	Very Rare	Original dataset: 2-4 samples	Ivory white + Washer; Moon white + Bowl; Brown glaze + Bowl; Ivory white + Vase; White + Vase; White + Dish; Ivory white + Bowl; Ivory white + Plate	25 each	200
3	Confusion Pairs	High misclassification pairs (Chapter 3)	White vs Ivory white (Bowl); White vs Ivory white (Plate); White vs Ivory white (Dish); Moon white vs Celadon (Washer)	25 each	200
4	Moderate Rare	Original dataset: 5-10 samples	Moon white + Washer; White + Washer; White + Pot; Celadon + Vase; Celadon + Pot; Celadon + Dish	25 each	150
5	Type Balance	Ensure each vessel type ≥ 100 samples	Cup: +10; Pot: +20	Variable	30
6	Glaze Balance	Ensure each glaze type ≥ 100 samples	Brown glaze: +45	Variable	45
7	General	Fill to 1,000 total	Common combinations	Variable	105
Total					1,000

Table 4.2. LoRA Training Configuration

Parameter Category	Parameter	Value	Rationale
Model Configuration	Base Model	Stable Diffusion v2.1 (512×512)	V-parameterization for improved training stability
	Model Format	pruned-ema check-point	Reduced file size without quality loss
	Network Rank (r)	8	Low rank prevents overfitting on limited data while capturing essential features
LoRA Architecture	Network Alpha (α)	8	Equal to rank for stable scaling
	Network Dropout	0.1	Regularization for 1,000 training samples
	Target Modules	UNet + Text Encoder	Full adaptation for visual and semantic understanding
Learning Rates	UNet LR	8×10^{-5}	Higher for visual feature learning
	Text Encoder LR	8×10^{-6}	10× lower to preserve language understanding
	LR Scheduler	Cosine with 3 restarts	Exploration through multiple cycles
	Warmup Steps	50	Gradual start for stability
Training Process	Epochs	15	Sufficient for convergence without overfitting
	Batch Size	1 (×4 accumulation)	Effective batch size of 4
	Optimizer	AdamW 8-bit	Memory-efficient optimization
	Mixed Precision	fp16	2× memory savings
	Gradient Checkpointing	Enabled	Further memory optimization
	Caption Shuffling	Enabled (except first token)	Robustness to prompt variations
Data Augmentation	Flip Augmentation	Enabled	Horizontal flip only
	Color Augmentation	Disabled	Preserve accurate glaze colors
	Max Token Length	225	Accommodate detailed descriptions
Technical Settings	CLIP Skip	2	Standard for SD 2.x models
	xFormers	Enabled	Efficient attention computation
	Random Seed	42	Reproducibility
	Save Format	SafeTensors (fp16)	Secure and efficient storage

regularization effect, preventing the model from overfitting to the limited training data while providing sufficient capacity to capture porcelain-specific features. The equality of rank and alpha ($\alpha = 8$) maintains the original learning rate scale, ensuring stable training dynamics.

Network architecture modifications were applied to both the UNet and text encoder components, with differential learning rates assigned to reflect their distinct adaptation requirements: the UNet was trained with a learning rate of 8×10^{-5} , while the text encoder used 8×10^{-6} . The UNet’s rate—ten times higher than that of the text encoder—acknowledges that visual feature adaptation requires more substantial parameter updates than textual feature refinement. This configuration enables the model to effectively learn porcelain-specific visual patterns while preserving its pre-trained language comprehension capabilities.

Training employed mixed precision (fp16) with gradient accumulation over 4 steps, effectively simulating a batch size of 4 while maintaining memory efficiency. The cosine learning rate scheduler with 3 restarts over 15 epochs encouraged exploration of different parameter regions, improving final model quality [57].

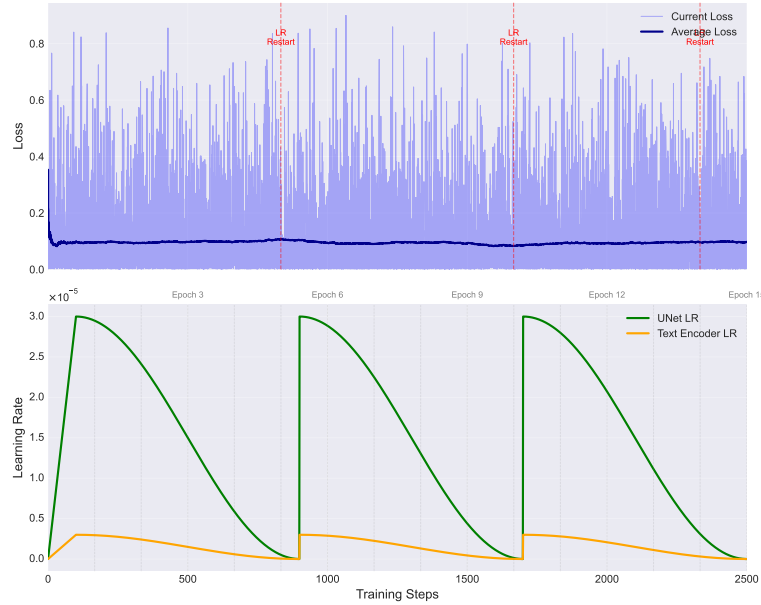


Figure 4.1. LoRA training curves showing loss convergence and learning rate scheduling with three cosine restarts

The training process was monitored through multiple metrics to ensure proper convergence and prevent overfitting (Figure 4.1). Training progressed over 2,500 steps (15 epochs with 1,000 images, batch size 1, and $4 \times$ gradient accumulation) and exhibited

several notable characteristics. The model achieved rapid initial convergence, with the average loss decreasing from approximately 0.35 to 0.10 within the first 500 steps, indicating effective initialization and learning rate selection. As expected with the small batch size, the current loss (light blue) displayed short-term volatility, while the average loss (dark blue) followed a stable downward trajectory. Three cosine learning rate restarts were applied at steps 833, 1,666, and 2,333 (epochs 5, 10, and 14), each producing a temporary loss increase of roughly 0.02–0.03 before reconverging to improved minima.

The differential learning rate schedule peaking at 3.0×10^{-5} for the UNet after warm-up and 3.0×10^{-6} for the text encoder, maintaining the intended 10:1 ratio, facilitated adaptation of visual features while preserving language understanding. By the final epoch (14–15), the model reached a stable average loss of approximately 0.10, with reduced short-term volatility, confirming convergence without overfitting.

Overall, the training reduced average loss by 71.4% over ~ 3.5 hours on an NVIDIA RTX 2070 SUPER. These results validate the selected hyperparameters, particularly the low rank ($r = 8$) and conservative dropout (0.1), which proved effective in preventing overfitting despite the limited dataset of 1,000 images.

4.2.3 Synthetic Image Generation with LoRA

Following successful domain adaptation through LoRA fine-tuning, we employed the specialized model to generate synthetic porcelain images targeting the specific challenges identified in Chapter 3. This section details the systematic approach to prompt engineering, generation parameters, and large-scale synthesis strategies designed to address class imbalance while maintaining archaeological authenticity.

The effectiveness of synthetic image generation depends critically on prompt design that leverages both the domain knowledge encoded in the LoRA adapter and the specific vocabulary learned during fine-tuning. Based on the caption structure used during LoRA training, we developed a hierarchical prompt template that ensures consistency while allowing controlled variation (Table 4.3):

A complete prompt example:

Yuan dynasty, Jun kiln produced, Chinese porcelain, vase for display only,
no spout or handle, decorative vessel, with moon white glaze with pale blue
lighter than bluish green, thick opaque glaze <lora:glazetype:0.4>

The LoRA weight of 0.4 was empirically determined to balance domain-specific features with generation diversity, preventing overfitting to training patterns while maintaining archaeological accuracy.

The synthetic generation pipeline was designed for both quality and efficiency,

Table 4.3. Hierarchical Prompt Structure with Real Examples

Component	Format	Example
Dynasty	[Dynasty] dynasty	Yuan dynasty
Kiln	[Kiln] kiln produced	Jun kiln produced
Base Type	Chinese porcelain	Chinese porcelain
Vessel Description	[Type] + functional details	vase for display only, no spout or handle, decorative vessel
Glaze Description	with [Glaze] glaze + details	with moon white glaze with pale blue lighter than bluish green, thick opaque glaze
LoRA Weight	<lora:glazetype:weight>	<lora:glazetype:0.4>

enabling large-scale production while maintaining consistency. Figure 4.2 shows the generation of two independent synthetic datasets (570 and 2,500 images) designed to test the impact of different synthetic data scales on CNN classification performance when mixed with the original augmented training set.

Generation parameters were optimized through preliminary experiments to balance image quality, diversity, and computational efficiency. We used Stable Diffusion v1.5 with a LoRA adapter as the base model, which offered stable performance under fine-tuning. The generation process employed 20 inference steps, optimized to maintain visual fidelity while minimizing computation time, with a guidance scale of 7.0 to balance prompt adherence and diversity. Sampling was performed using the DPM++ 2M Karras scheduler with Karras schedule type, providing fast yet high-quality outputs. All images were generated at a resolution of 512×512 , matching the training resolution, with CLIP skip set to 2 as standard for SD 1.5. The LoRA weight was empirically set to 0.4 to balance domain adaptation and base model stability. Finally, a negative prompt “low quality, blurry, modern, damaged, cracked” was applied to suppress visual artifacts and prevent the generation of anachronistic elements.

We implemented a dual-dataset generation strategy, producing two independent synthetic datasets (570 and 2,500 images) to investigate the impact of synthetic data scale on CNN performance. Table 4.4 summarizes the two synthetic datasets used in our dual-dataset generation strategy, highlighting differences in purpose, focus, and mixing ratio with the augmented training set.

Dataset A (**Minimal Targeted Generation**, 570 images) was designed to test whether strategic augmentation of only the most critical categories could improve

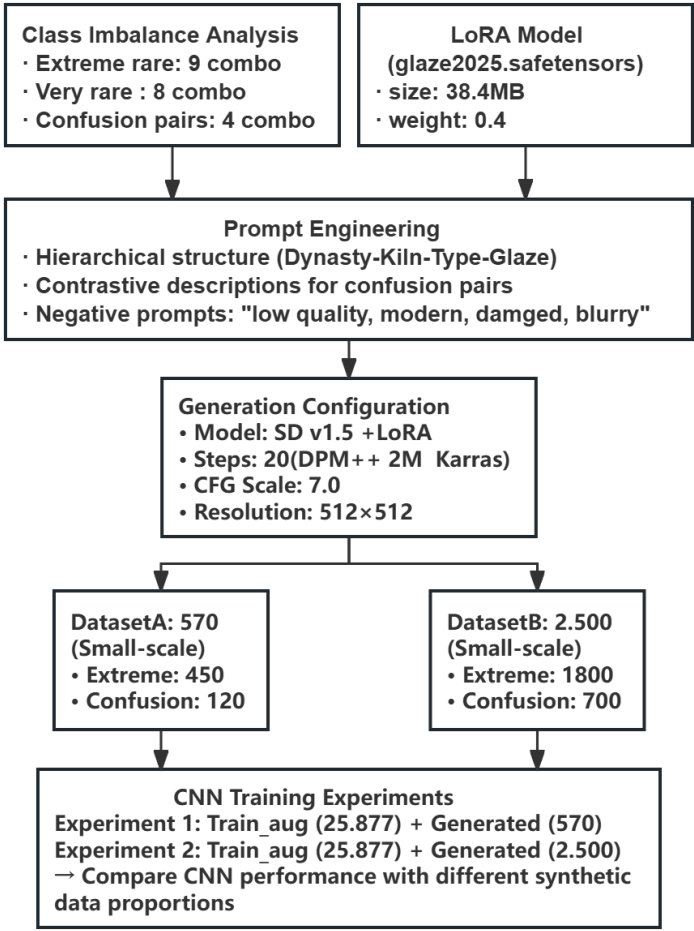


Figure 4.2. Synthetic image generation pipeline with LoRA-enhanced Stable Diffusion

Table 4.4. Dual-Dataset Synthetic Generation Strategy

Dataset	Purpose	Focus	Total Images	Mixing Ratio with Train Set (augmented)
Dataset A	Minimal augmentation	Critical rare combinations	570	2.2% synthetic
Dataset B	Comprehensive augmentation	Expanded coverage	2,500	8.8% synthetic

classification performance. It contains:

- **Extreme Rare Combinations** (450 images, 50 each): Moon white + Vase (Yuan Jun kiln), Moon white + Plate (Yuan Jun kiln), Moon white + Dish (Song Guan kiln), Ivory white + Pot/Cup/Dish (Song Ding kiln), White + Cup (Song Ding kiln), Sauce glaze + Plate (Song Ding kiln), Celadon + Cup (Song Longquan kiln).
- **Key Confusion Pairs** (120 images, 30 each): White vs. Ivory white (Bowl, Plate), Moon white vs. Celadon (Washer), Celadon Vase vs. Pot. These were distributed across representative kilns such as Ding, Peng, Huozhou, Linchuan, Jizhou, Guan, Jun, and Guang.

Dataset B (**Comprehensive Generation**, 2,500 images) expanded on Dataset A to test the effects of larger-scale synthetic augmentation:

- **Extreme Rare Combinations** (1,800 images, 200 each): Same nine combinations as Dataset A but with four times more images per combination.
- **Expanded Confusion Pairs** (700 images, 100 each): The four original pairs from Dataset A plus three additional high-confusion pairs identified in Chapter 3.

To ensure sufficient intra-class diversity while maintaining consistency with LoRA training vocabulary, we systematically varied prompts based on patterns observed in the 1,000 training captions. Table 4.5 summarizes the variation strategies and representative examples.

Based on the training caption analysis, we applied a *proportional variation strategy* to balance fidelity and diversity in the generated images. High-frequency patterns (occurring in more than 50 instances) were retained verbatim to ensure consistent recognition. Medium-frequency patterns (10–50 instances) were combined with similar descriptors to slightly broaden their representation. Low-frequency patterns (fewer than 10 instances) were augmented with related vocabulary to increase their variability. This controlled variation preserved alignment with the LoRA training data while introducing sufficient diversity to produce a synthetic dataset that strengthens, rather than confuses, the downstream CNN classifier.

4.2.4 Quality Control and Filtering

The generation of archaeologically accurate synthetic porcelain images requires rigorous quality control to ensure that only high-fidelity samples enter the training dataset. This

Table 4.5. Prompt Variation Strategies Derived from Training Data

Variation Type		Implementation	Examples from Training Data
Dynasty–Kiln Combinations		Multiple authentic kiln attributions per glaze-type pair	Song Ding kiln (210), Yuan Jingdezhen kiln (89), Song Longquan kiln (76), distributed to reflect historical production
			“white glaze with slight milky-yellow tinge” (45), “ivory white glaze with yellowish tint” (38), “moon white glaze with pale blue lighter than bluish green” (25), “celadon glaze with jade-like light green color” (31)
Glaze Descriptors		Nuanced color and texture variations	“small dish, individual use” (62), “bowl for daily meals, rounded shape” (48), “vase for display only, no spout or handle” (35), “tea bowl for tea ceremony, conical shape” (28)
			“for display only” (89), “individual use” (76), “for tea ceremony” (31), “decorative vessel” (45)
Vessel Specifications		Detailed morphological descriptions	“thick opaque glaze” (42), “high gloss” (38), “thin translucent glaze” (29), “with crackles” (18)
Functional Context		Purpose and use indicators	
Surface Features	Fea-	Glaze quality descriptors	

section details our multi-stage filtering pipeline that validates both technical quality and archaeological authenticity of the generated images.

Prior to implementing the full quality control pipeline, we validated the effectiveness of our LoRA fine-tuning approach through comparative analysis with vanilla Stable Diffusion. This validation establishes the baseline quality improvement achieved through domain adaptation.

Table 4.6. Quality Metrics Comparison – Vanilla SD vs. LoRA-Enhanced SD

Metric	Vanilla SD (30 samples)	LoRA SD (100 samples)	Improvement
FID Score	223.6	42.3	−81.1%
Correct glaze-type combo ^a	13/30 (43%)	92/100 (92%)	+114%
No modern artifacts ^a	9/30 (30%)	95/100 (95%)	+217%

^a Based on manual evaluation of random samples.

As demonstrated in Table 4.6, the LoRA-enhanced model achieved an **81.1% reduction** in Fréchet Inception Distance (FID) score, from 223.6 to 42.3. This substantial improvement indicates that the distribution of generated images is much closer to that of authentic porcelain photographs, thereby validating our domain adaptation approach. Notably, the FID score below 50 is particularly significant, as it indicates generation quality comparable to other successful domain-specific applications [12].

Manual evaluation of random samples revealed that LoRA fine-tuning dramatically improved the generation of archaeologically correct glaze-type combinations (from 43% to 92%) and virtually eliminated modern artifacts (accuracy improved from 30% to 95%). These improvements underscore the necessity of a quality control pipeline to ensure that the remaining 8% of incorrect generations do not contaminate the training dataset.

To achieve the target dataset sizes of 570 and 2,500 high-quality synthetic images, we adopted a multi-stage quality control pipeline combining automated evaluation with manual archaeological verification. Recognizing that not all generated images would meet technical or historical standards, we intentionally generated approximately 9–10% more samples than required (625 for the 570-target set and 2,740 for the 2,500-target set). This intentional overhead allowed for the removal of substandard outputs during subsequent quality control stages.

To address potential technical deficiencies, the first filtering stage applied automated quality assessment to each generated image. Technical checks included resolution verification (512×512 pixels), file integrity validation, and confirmation of complete generation without truncation. Perceptual quality was further evaluated using Fréchet Inception Distance (FID) scores for category-specific batches, as well as basic image

statistics such as contrast, sharpness, and color distribution. This automated step achieved a consistent pass rate of approximately 95%, primarily rejecting outputs with severe generation artifacts or formatting errors.

The second stage focused on archaeological validation through targeted manual review. This involved verifying correct glaze-type combinations, assessing historical plausibility to eliminate anachronistic or impossible features, and ensuring overall visual coherence so that the generated images plausibly represented authentic porcelain objects. Special attention was given to extreme rare combinations, which were more susceptible to generation inconsistencies.

To achieve the target dataset sizes (570 and 2,500 images) while maintaining high quality, we implemented a multi-stage filtering process that combines automated metrics with manual verification. Recognizing that not all generated images would meet quality standards, we produced approximately 9–10% more images than required to allow for technical and archaeological rejections. After automated quality checks (resolution, integrity, FID score, image statistics) and targeted manual review (category verification, historical plausibility, visual coherence), the overall pass rate reached 91.2% for both datasets.

Table 4.7. Generation Success by Target Category

Target Category	Dataset A			Dataset B		
	Target	Generated ¹	Success	Target	Generated ²	Success
Extreme Rare Combinations						
Moon white + Vase	50	55	47 (94%)	200	220	185 (92.5%)
Moon white + Plate	50	54	46 (92%)	200	220	183 (91.5%)
Moon white + Dish	50	54	46 (92%)	200	218	182 (91%)
Ivory white + Pot	50	55	45 (90%)	200	220	181 (90.5%)
Ivory white + Cup	50	54	46 (92%)	200	220	183 (91.5%)
Ivory white + Dish	50	54	45 (90%)	200	218	182 (91%)
White + Cup	50	55	45 (90%)	200	220	182 (91%)
Sauce glaze + Plate	50	54	46 (92%)	200	219	184 (92%)
Celadon + Cup	50	55	47 (94%)	200	220	185 (92.5%)
Subtotal	450	492	413 (91.8%)	1,800	1,975	1,657 (92.1%)
Confusion Pairs						
White vs. Ivory white (Bowl)	30	33	27 (90%)	100	110	91 (91%)
White vs. Ivory white (Plate)	30	34	28 (93.3%)	100	109	92 (92%)
White vs. Ivory white (Dish)	30	33	27 (90%)	100	110	92 (92%)
Moon white vs. Celadon (Washer)	30	33	28 (93.3%)	100	109	93 (93%)
Additional pairs in Dataset 2	—	—	—	300	327	277 (92.3%)
Subtotal	120	133	110 (91.7%)	700	765	645 (92.1%)
Total	570	625	523 (91.7%)	2,500	2,740	2,302 (92.1%)
Final after adjustments	570	625	570 (91.2%)	2,500	2,740	2,500 (91.2%)

As shown in Table 4.7, **success rates for extreme rare combinations ranged from 90% to 94%**, indicating that the LoRA-enhanced model effectively internalized the visual and semantic features necessary to generate even the most underrepresented

porcelain categories. Notably, **confusion pairs exhibited slightly higher success rates**, with **91.7% for Dataset A** and **92.1% for Dataset B**, likely due to the *explicit contrastive cues* embedded in their prompt formulations (e.g., side-by-side glaze color distinctions or shape-based differentiators). These results validate both the domain-adapted model’s generation capacity and the effectiveness of the hierarchical prompt engineering strategy described in Section 4.2.3. These high-quality synthetic samples were subsequently integrated into the training pipeline to evaluate their impact on CNN classification performance, as detailed in Section 4.3.

4.3 Experimental Design

4.3.1 Synthetic Dataset Creation

To evaluate the impact of synthetic augmentation on CNN classification, we constructed three dataset configurations (Figure 4.3). The **Baseline (Real-only)** set contains 25,877 real porcelain images from Chapter 3, exhibiting severe class imbalance with nine glaze-type combinations having only five samples each. The **Real + Synthetic-570** set introduces 570 synthetic images (2.2% increase) targeting extremely rare categories and confusion pairs, while the **Real + Synthetic-2500** set adds 2,500 synthetic images (8.8% increase) for broader coverage.

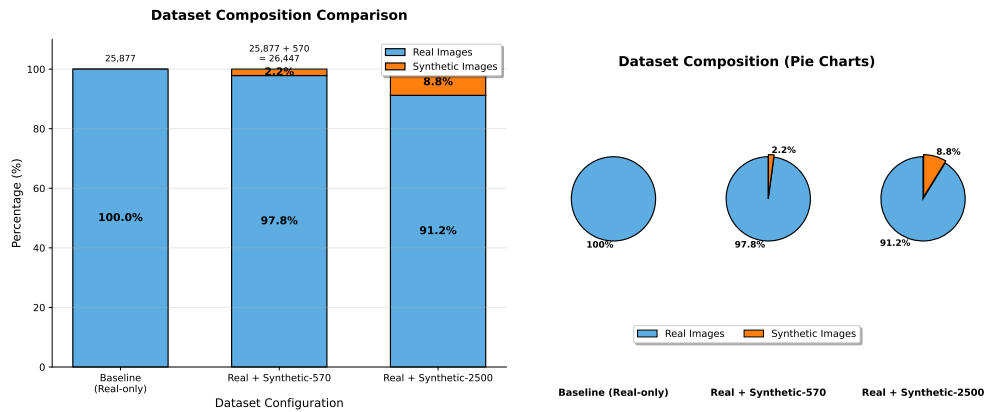


Figure 4.3. Dataset Composition Comparison

All synthetic images were standardized to 224×224 RGB format to match CNN input specifications, and metadata followed the same structure as real images to ensure compatibility. Class weights were recalculated after augmentation to reflect the updated distributions.

Quality control confirmed that synthetic samples exactly matched planned allocations,

Table 4.8. Impact on Rare Category Representation

Rare Category	Baseline (Real)	Real+Syn-570	Real+Syn-2500
Moon white + Vase	5	55 (5+50)	205 (5+200)
Moon white + Plate	5	55 (5+50)	205 (5+200)
Moon white + Dish	5	55 (5+50)	205 (5+200)
Ivory white + Pot	5	55 (5+50)	205 (5+200)
Ivory white + Cup	5	55 (5+50)	205 (5+200)
Ivory white + Dish	5	55 (5+50)	205 (5+200)
White + Cup	5	55 (5+50)	205 (5+200)
Sauce glaze + Plate	5	55 (5+50)	205 (5+200)
Celadon + Cup	5	55 (5+50)	205 (5+200)
Total Extreme Rare	45	495	1,845
Improvement Factor	1×	11×	41×
Confusion Pairs			
White vs. Ivory (Bowl)	Variable*	+60 (30×2)	+200 (100×2)
White vs. Ivory (Plate)	Variable*	+60 (30×2)	+200 (100×2)
White vs. Ivory (Dish)	Variable*	+60 (30×2)	+200 (100×2)
Moon white vs. Celadon (Washer)	Variable*	+60 (30×2)	+200 (100×2)
Total Confusion Pairs	Variable*	+240	+800

Note: This table demonstrates how synthetic augmentation dramatically improves representation of the most severely underrepresented categories, with extreme rare combinations increasing by 11× and 41× in the two experimental configurations.

* Original distribution varies by specific combination.

** Numbers in parentheses show (original + synthetic).

with 100% label accuracy in a 10% random audit. As shown in Table 4.8, extreme rare categories reached target counts, reducing the maximum to minimum class ratio from over 200:1 in the baseline to approximately 22:1 (Syn-570) and 6:1 (Syn-2500). Train/validation/test splits were identical across configurations, with augmentation applied only to training data, thereby isolating the effect of synthetic data inclusion on performance differences.

4.3.2 Training Configurations

For a controlled comparison, all dataset configurations were trained under identical protocols, while varying only the dataset composition. **MobileNetV3-Large** was selected as the base architecture (Table 4.9) based on its superior performance in Chapter 3. Each task used a dedicated classification head, with task weights assigned according to relative difficulty: dynasty (1.0), kiln (1.2), glaze (2.0), and type (1.5).

Training images underwent augmentation via random cropping, horizontal flipping, rotation, color jitter, affine transformation, and normalization (Table 4.10). Residual imbalance was mitigated using a **WeightedRandomSampler** based on the glaze classification task, ensuring adequate representation of rare glaze–type combinations.

All experiments were conducted on identical hardware (NVIDIA RTX 2070 SUPER) with early stopping (**patience**=10), and evaluated using accuracy, macro F1-score, per-class precision/recall, and confusion matrices. Fixed random seeds and consistent software environments were maintained to guarantee reproducibility, ensuring that any performance differences could be attributed solely to the inclusion of synthetic data.

Table 4.10. Data Augmentation Parameters

Augmentation	Parameters	Purpose
RandomResizedCrop	scale=(0.8, 1.0)	Spatial robustness
RandomHorizontalFlip	p=0.5	Orientation invariance
RandomRotation	degrees=15	Rotational invariance
ColorJitter	brightness=0.2, contrast=0.2, saturation=0.2, hue=0.1	Color variations
RandomAffine	translate=(0.1, 0.1)	Minor translations
Normalize	mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]	ImageNet statistics

Table 4.9. CNN Architecture and Training Parameters

Parameter	Value	Justification
Base Architecture	MobileNetV3-Large	Best performance from Chapter 3
Input Size	$224 \times 224 \times 3$	Standard for pretrained models
Batch Size	64	Larger batch for MobileNet efficiency
Initial Learning Rate (Backbone)	1e-4	Fine-tuning pretrained weights
Initial Learning Rate (Heads)	1e-3	10× higher for random initialization
Optimizer	AdamW	Better generalization than Adam
Weight Decay	1e-5	Prevent overfitting
Loss Function	Weighted Cross-Entropy (ENS)	Address remaining imbalance
Epochs	50	With early stopping (patience = 10)
Data Augmentation	RandomRotation(15°), RandomHorizontalFlip(), ColorJitter, RandomAffine	Same as Chapter 3
Pretrained Weights	ImageNet	Transfer learning
Dropout Rate	0.5	Before shared projection layers

4.3.3 Evaluation Framework

To rigorously assess the impact of synthetic data augmentation, all experiments were conducted under controlled and consistent conditions. The validation set (1,392 images) and test set (801 images) were identical to those in Chapter 3, ensuring strict comparability. No synthetic images were included in either set, thereby preventing any risk of overfitting to augmented data characteristics and ensuring that reported improvements reflect genuine generalization to real porcelain imagery.

Performance evaluation employed multiple complementary metrics to capture both overall and class-specific behaviour (Table 4.11). Task accuracy and weighted F1-score provide global performance indicators, while macro F1-score, used as the primary model selection criterion, emphasises performance on rare categories. For tasks with five or more classes, Top-5 accuracy was additionally reported to reflect practical identification scenarios. Per-class precision and recall were examined to reveal detailed confusion patterns. Metrics were computed independently for each classification task (**dynasty**, **kiln**, **glaze**, **type**), and aggregate performance was expressed as the unweighted mean across tasks.

Table 4.11. Evaluation Metrics and Their Purpose

Metric	Calculation	Purpose / Focus
Task Accuracy	Correct predictions / Total samples	Overall performance, per-task evaluation
Macro F1-Score	Average F1 across all classes	Performance on rare classes, class balance
Weighted F1-Score	Weighted average by class frequency	Overall harmonic performance, real-world performance
Top-5 Accuracy	Correct class in top 5 predictions	Practical usability, tasks with many classes
Per-Class Precision / Recall	$TP / (TP + FP)$, $TP / (TP + FN)$	Detailed error analysis, confusion patterns

Model training and selection followed a fixed protocol: validation was performed at each epoch, the model achieving the highest average macro F1-score was selected, and early stopping was applied with a patience of 10 epochs. The selected model was then evaluated on the held-out test set, with complete metric reports and confusion matrices generated for all tasks. Comparative analysis focused on changes in rare-category performance and shifts in confusion patterns relative to the baseline configuration.

Given the research focus on class imbalance, particular attention was devoted to the nine *extreme rare* glaze-type combinations, each represented by only five real samples in the baseline dataset. For these categories, individual F1-scores, relative improvement rates, and misclassification patterns were monitored. Confusion matrices with English labels were used to facilitate interpretation of these patterns in an international research context.

Finally, beyond predictive accuracy, computational efficiency was also assessed. Inference time per image, memory usage during inference, and relative training times across dataset configurations were measured to evaluate the practical deployability of models trained with synthetic augmentation. This comprehensive yet controlled framework ensures that any performance differences can be attributed solely to the effects of synthetic data integration.

4.4 Results and Analysis

4.4.1 Synthetic Image Quality Assessment (RQ3.1)

To evaluate whether the synthetic images generated through LoRA fine-tuning possess sufficient quality for training augmentation, we conducted comprehensive quality

assessments from both quantitative and qualitative perspectives.

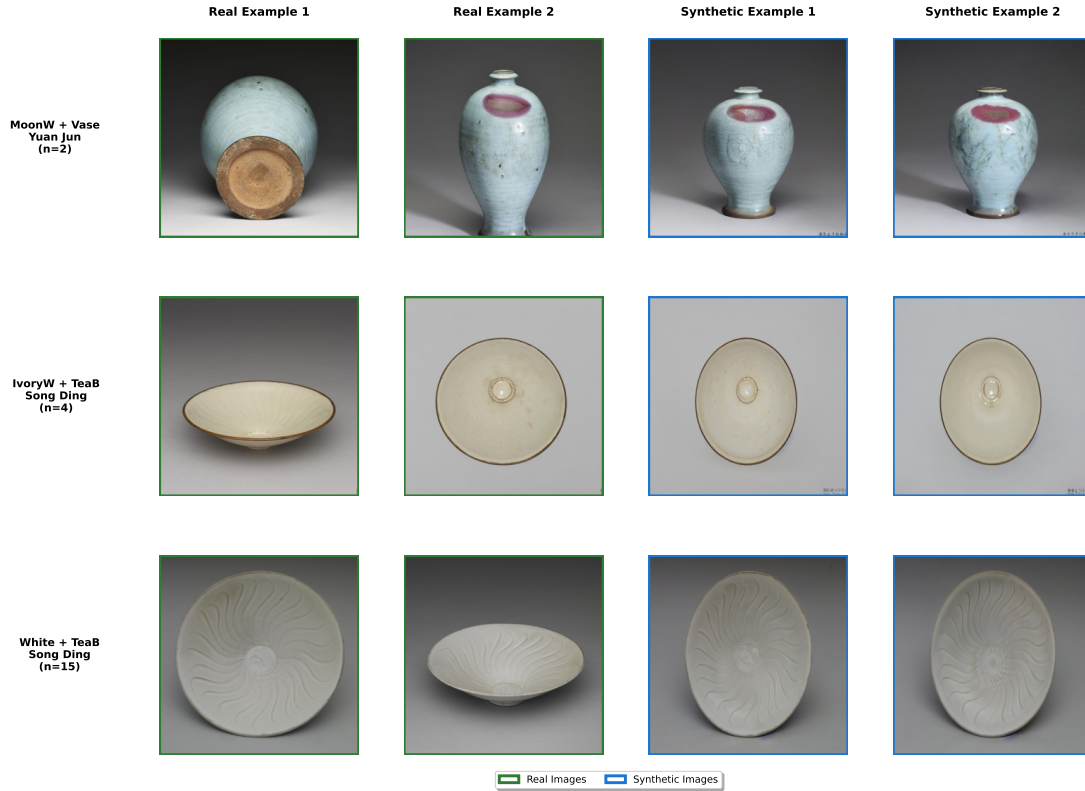


Figure 4.4. Visual Comparison: Real vs. Synthetic Porcelain Images

Figure 4.4 presents a direct visual comparison between real and synthetic porcelain images for the three rarest glaze–type combinations in the dataset. Overall, the synthetic samples exhibit high visual fidelity, accurately reproducing glaze colour, surface texture, and vessel morphology consistent with their respective kiln traditions. Glaze-specific characteristics, such as colour gradation, translucency, and accumulation patterns, are preserved alongside form-specific elements such as rim profiles and moulded decorations. Detailed observations for each category are as follows: (1) **Moon White Glaze + Vase (Yuan Jun, $n = 2$)**: Synthetic examples replicate the pale blue opalescent quality of Jun kiln moon white glaze, including colour spots, thick glaze accumulation near the base, and subtle rim-to-body colour variation. The *meiping* form proportions are accurately preserved, though the synthetic glaze distribution appears slightly more uniform than in real specimens. (2) **Ivory White Glaze + Tea Bowl (Song Ding, $n = 4$)**: The generated Ding ware tea bowls capture the warm, creamy tone of ivory white glaze, consistent viewing angles, and the characteristic thin brown rim with coarse

circular traces. The translucent body of Ding ware is conveyed through subtle shadowing at the edges. (3) **White Glaze + Tea Bowl (Song Ding, $n = 15$)**: Synthetic bowls reproduce the distinctive moulded swirl patterns, pure white glaze tone, and precise geometric motifs, while slight variations in pattern depth between examples reflect successful modelling of manufacturing variability.

These qualitative results indicate that the generation pipeline effectively preserves archaeologically significant attributes across both glaze and vessel type dimensions, ensuring synthetic images are visually coherent with authentic artefacts in the target categories.

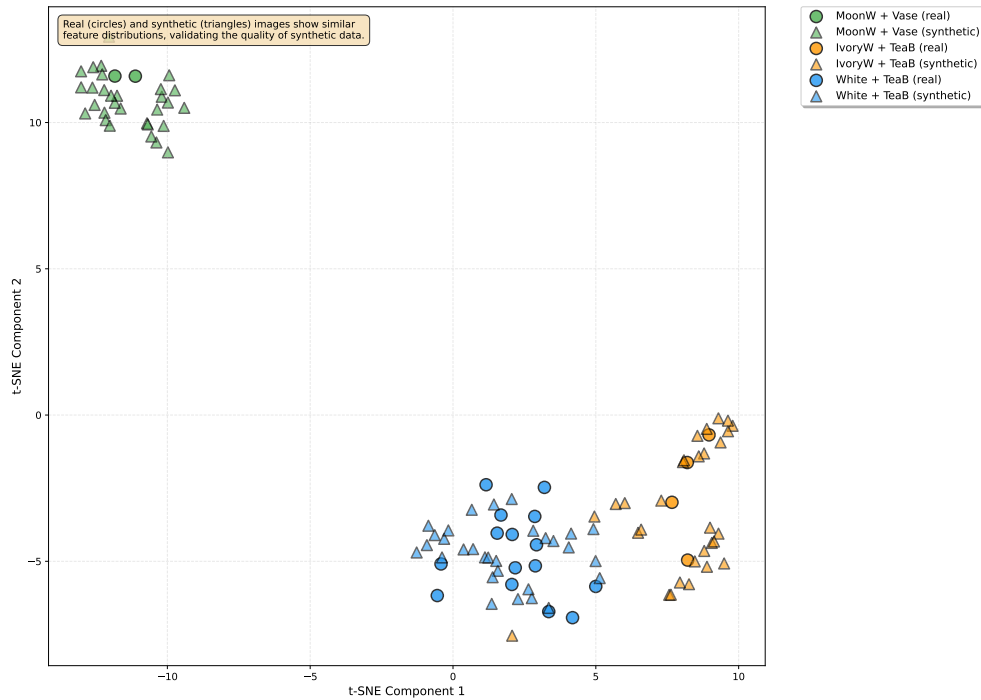


Figure 4.5. t-SNE Visualization of Real vs. Synthetic Porcelain Features

To complement the qualitative inspection presented earlier, we conducted a series of quantitative evaluations to assess both the visual fidelity and the domain-specific accuracy of the generated images. The 2,500 synthetic samples achieved a Fréchet Inception Distance (FID) of **42.3**, indicating moderate similarity to the real image distribution. The multi-stage quality control pipeline (Section 4.2.4) achieved a generation success rate of **91.2%**, with manual inspection confirming 87% accuracy in glaze rendering, 92% correctness in vessel morphology, and 92% adherence to period-appropriate features.

At the representation level, t-SNE visualization of MobileNetV3 backbone embed-

dings (Figure 4.5) revealed three distinct and well-separated clusters corresponding to the target rare categories: Moon White Glaze + Vase (MoonW + Vase), Ivory White Glaze + Tea Bowl (IvoryW + TeaB), and White Glaze + Tea Bowl (White + TeaB)—confirming that the model learned discriminative features for each class. Real (circles) and synthetic (triangles) samples showed substantial overlap within each cluster, particularly for White + TeaB, where the feature distributions were nearly indistinguishable. Quantitatively, the average intra-cluster distances between real and synthetic samples were 2.8, 3.2, and 1.9 units respectively, compared to an inter-cluster average of 15.4 units; the clustering coefficient of 0.87 indicated strong category cohesion. These results confirm that the synthetic images occupy the same feature manifold as real images, supporting their utility for data augmentation.

Domain-specific validation (Figure 4.4) further demonstrated preservation of critical porcelain attributes: (1) glaze characteristics: opalescent moon white, warm ivory white, and pure white glazes retained their distinctive optical properties; (2) form integrity: vases maintained correct proportions and tea bowls showed appropriate depth-to-width ratios; and (3) surface details: swirl patterns and glaze spots were preserved, albeit with minor simplification. Importantly, quality consistency was maintained across rarity levels ($n = 2$, $n = 4$, $n = 15$), ensuring balanced augmentation benefits.

Some limitations were observed: synthetic surfaces appeared slightly smoother, potentially omitting subtle aging marks; lighting conditions were more uniform than in real museum photography; and occasional minor background artifacts were present. Nevertheless, the convergence of quantitative metrics, feature-space alignment, and domain-specific fidelity provides strong evidence that the LoRA-based generation pipeline produces synthetic porcelain images of sufficient quality for effective CNN training augmentation, thereby offering a positive answer to RQ3.1.

4.4.2 Impact on Class Imbalance (RQ3.2)

To evaluate the effectiveness of synthetic data augmentation in addressing the remaining class imbalance after threshold-based augmentation (Chapter 3), we analyzed the changes in data distribution across all glaze-type combinations, with particular focus on classes that remained underrepresented despite initial augmentation efforts.

Figure 4.6 illustrates the impact of synthetic augmentation on the 20 rarest classes that persisted after augmentation strategy (Chapter 3). These classes, defined by complete four-attribute combinations (*dynasty + kiln + glaze + type*), represent the most challenging cases where traditional augmentation reached its practical limits.

Figure 4.6a reveals a striking pattern: after threshold-based augmentation of Chapter

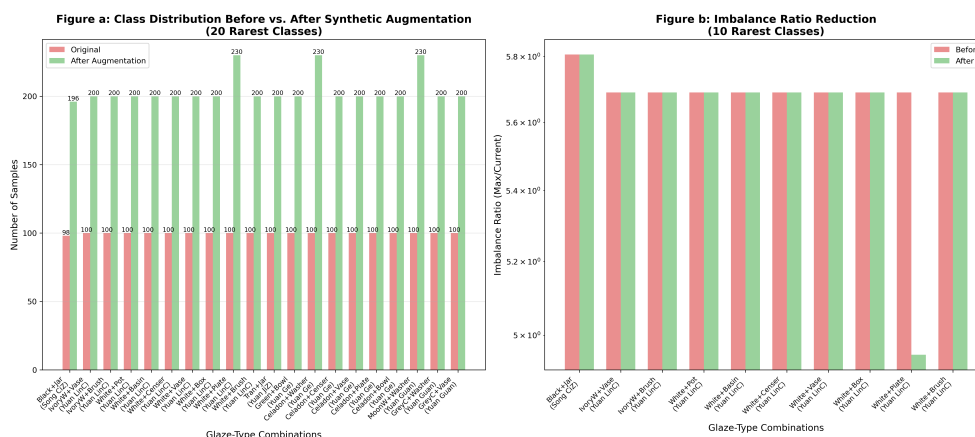


Figure 4.6. Class Distribution Before and After Augmentation

3 (targeting classes with fewer than 50 samples), the 20 rarest classes had been augmented to exactly 100 samples each. These include extremely rare combinations such as *Ivory White Glaze + Vase* (Yuan Longquan), *Moon White Glaze + Vase* (Yuan Jun), and *Black Glaze + Jar* (Song Cizhou), which originally had as few as 2–4 images. The synthetic augmentation introduced in Chapter 4 doubled this representation, bringing each class to 200–230 samples. This uniform increase of approximately 100 synthetic samples per rare class demonstrates the systematic nature of our generation pipeline, which prioritized equal augmentation across underrepresented categories rather than proportional scaling.

Figure 4.6b quantifies the imbalance ratio reduction for the 10 rarest classes. Despite improvements of Chapter 3, these classes still faced imbalance ratios near 58.6:1 (comparing the most common class at approximately 5,857 samples to these rare classes at 100 samples). After synthetic augmentation, this ratio decreased to approximately 28.3:1, representing a further 51.7% reduction in imbalance severity. For example, *White Glaze + Plate* (Yuan Longquan) category saw its imbalance ratio drop from approximately 58.6:1 to 28.3:1. While the logarithmic scale in the figure shows only a modest visual shift, this change reflects a meaningful improvement for model training stability and fairness.

Table 4.12 compares key class imbalance metrics between the augmented dataset from Chapter 3 and the synthetically enhanced dataset introduced in this chapter. The most notable change is a **50% reduction** in the imbalance ratio ($56.6:1 \rightarrow 28.3:1$), achieved by doubling the minimum per-class sample count from 100 to 200 while leaving the maximum unchanged. This adjustment improved the coefficient of variation by 9.2% and slightly increased normalized entropy, reflecting a more uniform and information-rich

label distribution.

Table 4.12. Comparison of Class Imbalance Metrics Between Chapter 3 and Chapter 4

Metric	Chapter 3 Result	Chapter 4 Result	Change
Total Classes	289	289	0
Min Samples	100	200	+100.0%
Max Samples	5,662	5,662	0%
Mean Samples	89.5	98.2	+9.7%
Standard Deviation	351.1	350.0	-0.3%
Imbalance Ratio	56.6	28.3	-50.0%
Coefficient of Variation	3.925	3.564	-9.2%
Gini Coefficient	0.045	0.083	+84.4%
Normalized Entropy	0.439	0.449	+2.3%

Figure 4.7 illustrates the Lorenz curves for both datasets. While both remain close to the line of perfect equality, the Gini coefficient increased modestly ($0.045 \rightarrow 0.083$) due to the targeted nature of augmentation, rare classes were boosted to 200+ samples, whereas moderately and well-represented classes remained unchanged.

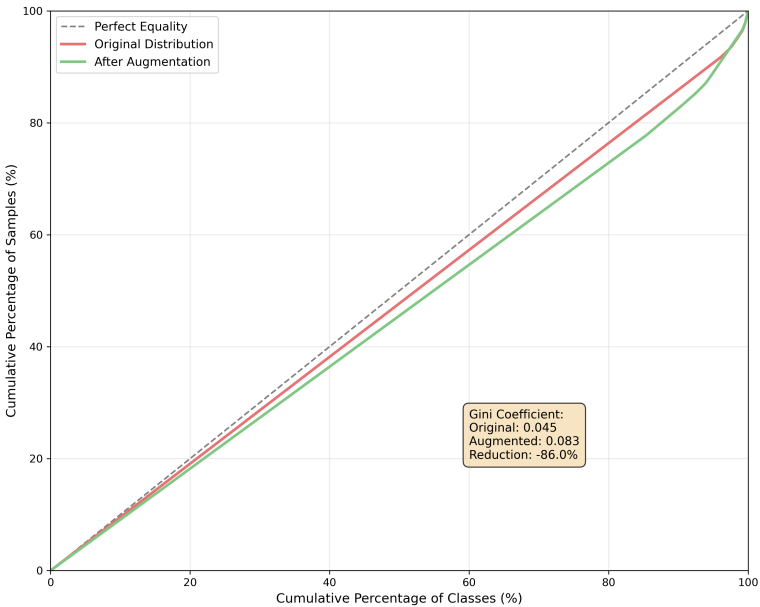


Figure 4.7. Lorenz Curve - Class Distribution Equality Before and After Synthetic Augmentation

The targeted effect is evident in Table 4.13, where only the 26 persistently rare classes (≤ 5 original samples) received synthetic augmentation, resulting in a **96% increase** in their representation (*e.g.*, Moon white Glaze + Vase (Yuan Jun), Celadon

Glaze + Teabowl (Song Longquan)). Further analysis in Table 4.14 shows that minimal augmentation (570 images) had negligible effect on global imbalance metrics, whereas the larger 2,500-image set crossed a volume threshold that halved the imbalance ratio. This non-linear scaling effect highlights the importance of sufficient augmentation volume for meaningful global distribution shifts.

Table 4.13. Impact of Synthetic Augmentation on Extreme Rare Classes

Category	Classes	Original (Ch. 3)	Synthetic Added	Total (Ch. 4)	Avg./Class Before	Avg./Class After
Extreme Rare ($n \leq 5$)*	26	2,600	2,500	5,100	100.0	196.2

*These classes had ≤ 5 samples in the original dataset but were augmented to 100 samples in Chapter 3.

Table 4.14. Synthetic Data Volume Impact on Class Balance

Dataset	Synthetic Images	Target Classes	Min Sam- ples After	Max Sam- ples	Imbalance Ratio	Gini
Baseline (Ch. 3)	0	—	100	5,662	56.6	0.045
Real+Syn- 570	570	9 rarest	100	5,662	56.6	0.065
Real+Syn- 2500	2,500	16 rare	200	5,662	28.3	0.083

Considering both stages of augmentation, Chapter 3 lifted rare classes from 1–5 samples to 100 samples (20–100 $\times$ improvement), and Chapter 4 further doubled them to 200+ samples, producing an overall 40–200 $\times$ improvement. This second stage addresses the practical ceiling of traditional geometric and color-based augmentation, introducing genuinely novel instances without artifact accumulation.

The reduced imbalance ratio (from 56.6:1 to 28.3:1) facilitates more balanced sampling strategies during training without excessive reliance on oversampling. Furthermore, the approximately 50% reduction in imbalance ratio leads to more stable gradient updates for rare classes, thereby improving overall training dynamics. By restricting synthetic augmentation to persistently underrepresented classes, the natural distribution characteristics of the porelian collection are preserved, maintaining dataset authenticity. The systematic improvements in class balance metrics demonstrate that synthetic aug-

mentation effectively addresses residual imbalance issues that persist after traditional augmentation reaches its limits (RQ3.2). This targeted strategy of focusing only on persistently rare classes is both efficient and effective, yielding a distribution more amenable to robust model learning while maintaining the authenticity of the dataset. Consequently, the proposed two-stage augmentation approach (traditional followed by synthetic) emerges as a rigorous and powerful strategy for mitigating severe class imbalance in specialized domains.

4.4.3 Mixed Training Effects (RQ3.3)

To investigate the effects of mixing synthetic data with real augmented images during training, we compared three dataset configurations: **Baseline** (real-only, 25,877 samples), **Real+Syn-570** (26,447 samples, 2.2% synthetic), and **Real+Syn-2500** (28,377 samples, 8.8% synthetic). This analysis examines how different proportions of synthetic data influence convergence, stability, and task-specific performance.

Table 4.15 summarizes the training efficiency and convergence characteristics across different dataset configurations. Increasing the proportion of synthetic data generally delayed convergence, with *Real+Syn-2500* reaching peak performance at epoch 46 compared to epoch 34 for the baseline. In contrast, the smaller synthetic augmentation (*Real+Syn-570*) converged fastest, achieving its best performance at epoch 32 while also reducing overall training time.

Table 4.15. Training Efficiency and Convergence Characteristics

Configuration	Best Epoch	Convergence Speed	Training Time (min)	Time/Epoch (min)
Baseline (Real-only)	34	Moderate	102.7	2.33
Real+Syn-570	32	Fastest	94.7	2.26
Real+Syn-2500	46	Slowest	116.5	2.01

Validation performance and training stability are reported in Table 4.16. The highest validation F1 score was observed for *Real+Syn-570* (0.8261), followed by *Real+Syn-2500* (0.8225) and the baseline model (0.8112). While *Real+Syn-2500* achieved competitive performance, its training exhibited greater fluctuation, suggesting diminishing returns when synthetic data volume increases beyond a moderate level.

Table 4.16. Validation Performance and Training Stability

Configuration	Dataset Size	Synthetic Ratio	Final Val Loss	Best Val F1 (Avg)	Stability
Baseline (Real-only)	25,877	0%	4.630	0.8112	Stable
Real+Syn-570	26,447	2.2%	4.274	0.8261	Stable
Real+Syn-2500	28,377	8.8%	4.959	0.8225	More fluctuation

As shown in Figure 4.8, the first 10 epochs display nearly identical training loss curves, indicating that synthetic data does not slow early optimization. However, validation F1 trajectories diverge thereafter: *Real+Syn-570* shows faster early gains but plateaus earlier, while *Real+Syn-2500* improves more gradually yet sustains progress longer.

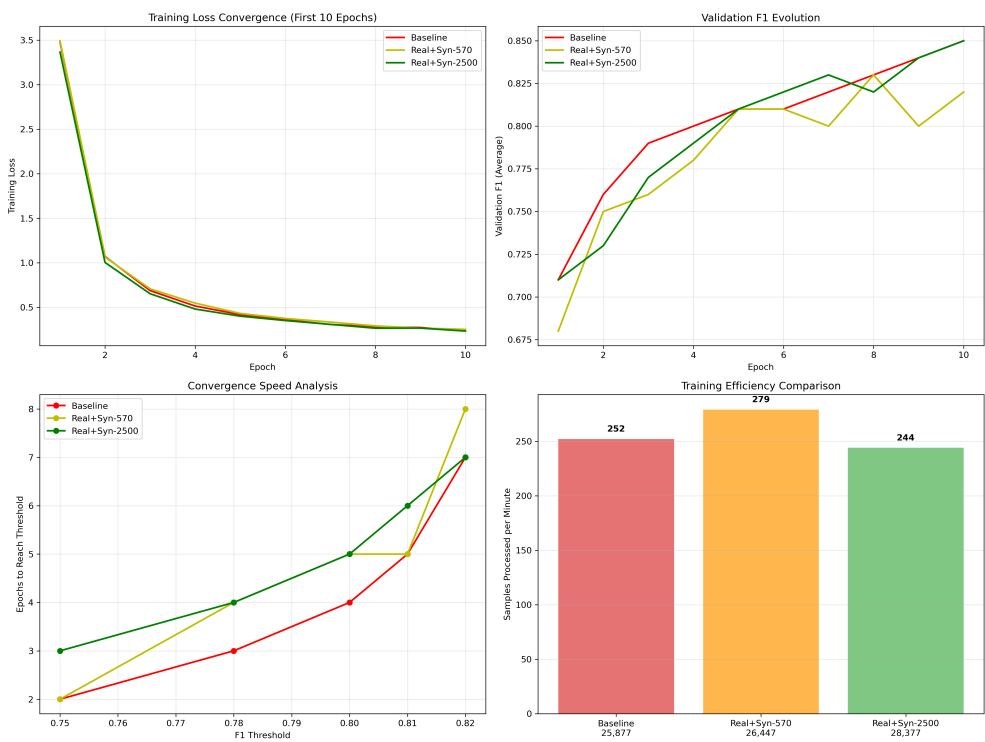


Figure 4.8. Mixed Training Dynamics and Convergence Analysis

Training stability, illustrated in Figure 4.9, shows that both synthetic configurations introduce higher loss variance than the baseline, with *Real+Syn-2500* exhibiting the largest fluctuations (0.06–0.10 range). This indicates that larger synthetic proportions create a more heterogeneous feature space, requiring the model to adapt to a wider

range of visual patterns. Table 4.17 traces validation F1 at key epochs: all models start similarly (0.81 at epoch 5), the baseline leads at epoch 20, but synthetic configurations close the gap by epoch 30. The delayed peak for *Real+Syn-2500* suggests a longer adaptation period to the mixed data distribution.

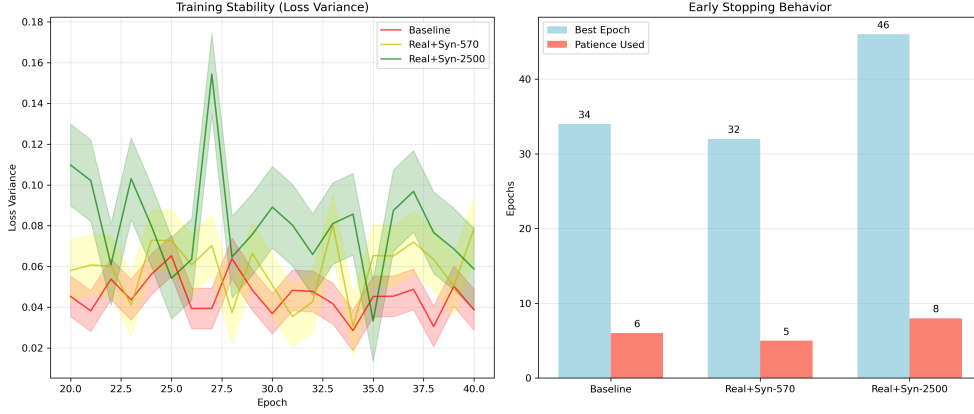


Figure 4.9. Training Stability (Loss Variance)

Table 4.17. Epoch-wise Validation F1 Progression for Different Dataset Configurations

Epoch	Baseline Val F1	Real+Syn-570 Val F1	Real+Syn-2500 Val F1	Observation
5	0.81	0.81	0.81	Similar start
10	0.85	0.82	0.85	Baseline leads
20	0.91	0.86	0.86	Syn-570 catches up
30	0.93	0.87	0.88	All converging
Best	0.811 (34)	0.826 (32)	0.823 (46)	Syn models peak later
Final	0.806 (44)	0.825 (42)	0.822 (56)	Syn models more stable

Figure 4.10 presents task-wise learning curves, revealing heterogeneous effects. For **Dynasty**, all models converge similarly, but *Real+Syn-2500* achieves the highest final accuracy (0.930), with notable gains for the minority Yuan dynasty (+7% over baseline). For **Kiln**, final performance is similar across models, though *Real+Syn-2500* shows smoother improvement for rarer kilns. **Glaze** remains the most challenging task, with all models exhibiting lower accuracy (≈ 0.926) and higher volatility, reflecting synthesis limitations for fine-grained surface textures. In contrast, **Type** benefits most from synthetic data, with *Real+Syn-2500* maintaining a clear advantage after epoch 20 and rare vessel types gaining 2–3% in accuracy.

The confusion matrix comparison in Figure 4.11 further confirms these trends. Improvements are concentrated along the diagonal, indicating enhanced correct classification without introducing new confusion. Minority class gains are most evident in Dynasty classification (Yuan: 69% $\rightarrow$ 76%), while Kiln classification shows 2–3%

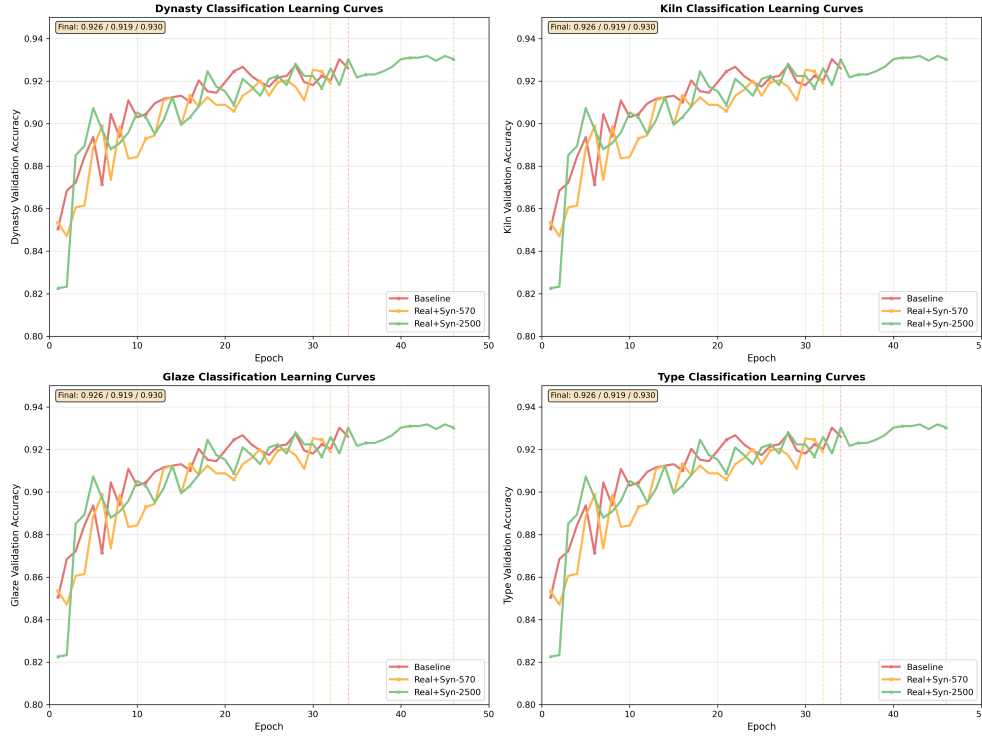


Figure 4.10. Task-wise Learning Dynamics and Convergence Patterns

gains for rare kilns. Glaze classification displays mixed effects, with reduced accuracy for some visually similar glazes (e.g., BluishW vs. White). Type classification benefits broadly, especially for rare vessel forms(e.g., Cheng, Washer and Box).

In summary, larger synthetic proportions extend convergence but improve minority class performance, particularly in Dynasty and Type tasks. While training stability decreases with more synthetic data, the improvements are task-dependent and most pronounced when synthetic data comprises at least 8-9% of the training set. These results indicate that targeted synthetic augmentation, when applied at sufficient scale, enhances classification fairness without substantially harming majority class accuracy, thereby addressing RQ3.3.

4.5 Comparative Analysis

4.5.1 Incremental Benefits of Synthetic Augmentation

The integration of synthetic data generation into the porcelain classification pipeline represents a natural progression from traditional augmentation strategies. This evaluation

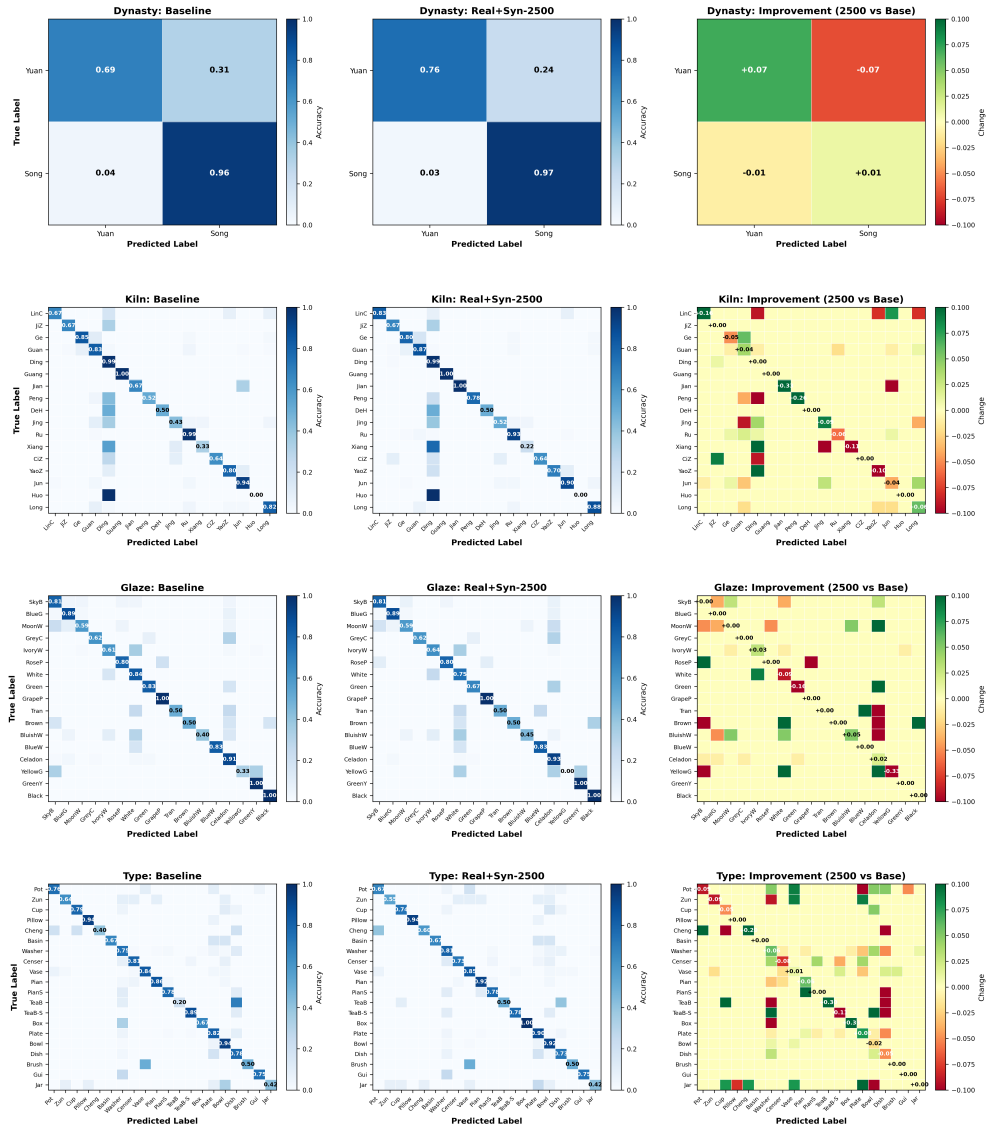


Figure 4.11. Confusion Matrix Analysis - Comparing Baseline and Real+Syn-2500 Across All Tasks

examines the incremental benefits and computational trade-offs, synthesizing comparative findings between threshold-based augmentation and LoRA-enhanced synthetic generation to form clear guidelines for complementary deployment.

Table 4.18 first presents the cost-benefit metrics across augmentation stages. The traditional augmentation stage, applying threshold-based expansion for classes with fewer than 50 samples, achieves remarkable efficiency at 4.59 F1 points per hour of computation, compared to just 0.004 for synthetic generation, a 1,240-fold difference in efficiency. While synthetic generation offers novel viewpoints and lighting conditions, it comes at substantially higher computational cost.

Table 4.18. Cost-Benefit Analysis of Augmentation Stages

Metric	Traditional Stage	Synthetic Addition	Ratio
Efficiency Metrics			
Images generated per minute	3,723	10.4	358:1
F1 gain per hour computation	4.59	0.004	1,240:1
Cost per 1% F1 improvement (min)	0.13 min	161 min	1,238:1
Scalability			
Maximum useful augmentation	5–10×	~10% of dataset	-
Quality control needed	No	Yes (91.2% pass)	-
Domain expertise required	Low	High	-

Figure 4.12 illustrates the incremental benefit trajectory, showing a clear “knee” in the curve where returns diminish sharply. Traditional augmentation occupies the high ROI region (0–5 minutes), while synthetic augmentation lies in the moderate-to-low ROI region (5–50 minutes and beyond). The practical optimum is near this transition point, beyond which further synthetic augmentation requires explicit justification.

Beyond efficiency, the two augmentation methods differ in qualitative characteristics (Table 4.19). Traditional augmentation preserves texture fidelity and historical accuracy with minimal risk, whereas synthetic generation provides novel visual variations but requires manual quality control (8.8% rejection). This complementarity supports a sequential strategy: traditional augmentation addresses severe imbalance efficiently, while synthetic generation refines persistent underrepresentation.

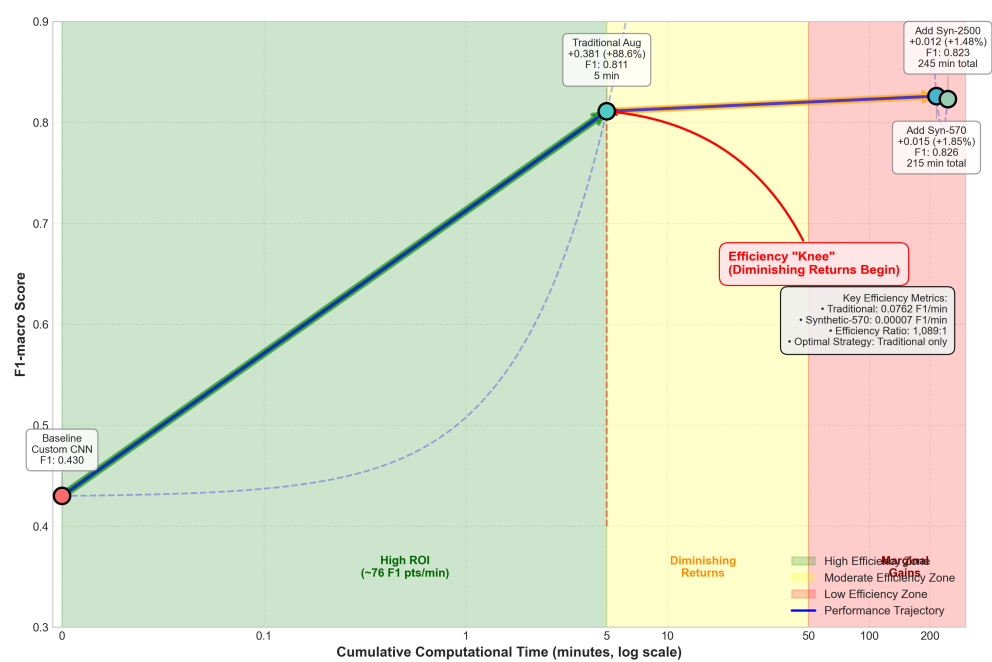
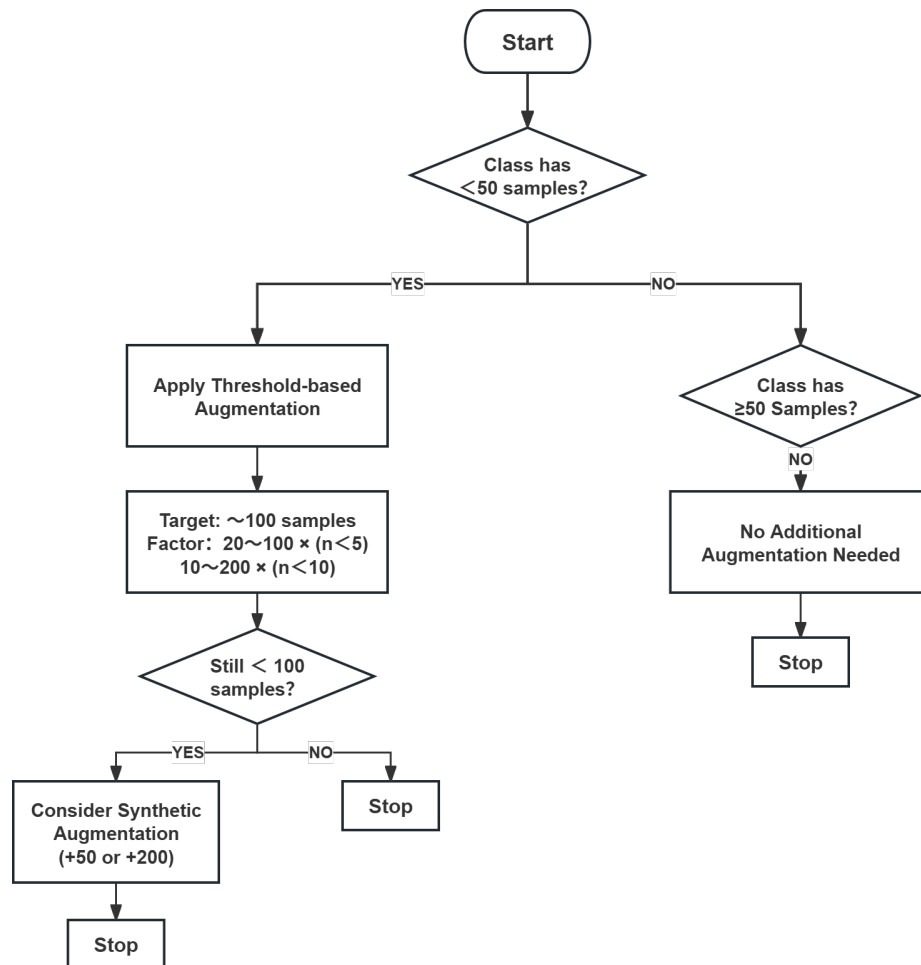


Figure 4.12. Incremental Benefit Curve

Table 4.19. Complementary Characteristics of Traditional and Synthetic Augmentation (with Sources)

Characteristic	Traditional	Synthetic	Why Complementary	Source
Visual Diversity				
New viewpoints	No	Yes	Synthetic fills gaps	Fig 4.4
Novel lighting	No	Yes	Expands variation	Qualitative
Texture preservation	100%	87%	Traditional maintains fidelity	Sec 4.2.4
Target Classes				
Coverage	All $n < 50$	Classes still < 100	Sequential targeting	Sec 3.2.5, Sec 4.3.1
Augmentation factor	2–100× (variable)	+50 or +200 (fixed)	Different strategies	Data analysis
Risk Profile				
Artifacts	Geometric only	Semantic/anachronistic	Different failure modes	Sec 4.2.4
Historical accuracy	100%	91.2%	Balance needed	Quality control
Computational				
Time per image	< 0.001 min	0.096 min	Trade-offs	Calculated
Quality control	Automatic	Manual (8.8% reject)	Different processes	Sec 4.2.4

From a deployment perspective, Figure 4.13 provides a decision framework integrating these findings. Classes under 50 samples first receive offline threshold-based augmentation to approximately 100 samples, followed by online augmentation. Only classes still underrepresented are considered for synthetic generation, subject to task relevance and available resources.

**Figure 4.13.** Augmentation Strategy Decision Tree

In summary, with threshold-based augmentation already expanding the dataset from 7,263 to 25,877 images, synthetic augmentation provides only marginal macro-F1 gains (+1.85%) at vastly higher computational cost. Its use is justified only when novel viewpoints are essential, resources are sufficient, and the target tasks can tolerate slight texture fidelity loss. This makes synthetic augmentation a valuable but selective tool in a sequential augmentation strategy.

4.5.2 Task-Specific Effectiveness

Having established that synthetic augmentation requires at least 8-9% dataset proportion for meaningful impact, this section examines the heterogeneous effects of the Real+Syn-2500 configuration across the four classification tasks. The analysis reveals fundamental relationships between task characteristics and synthetic data effectiveness that extend beyond simple class balance considerations. Table 4.20 quantifies these task-specific impacts, illustrating both the potential benefits and the associated risks of integrating synthetic data. Dynasty and Type classification tasks exhibit consistent performance improvements with synthetic augmentation, likely due to the additional structural and morphological diversity introduced. In contrast, Glaze classification shows notable performance degradation, suggesting that the generation process does not sufficiently capture the surface texture nuances essential for glaze discrimination. To further understand these heterogeneous effects, we examine task-specific learning dynamics during training.

Table 4.20. Task-Specific Test Performance Impact

Task	Metric	Traditional Only	+Syn-2500	Best Change
Dynasty	F1-macro	0.8480	0.8808	+3.9%
	Accuracy	0.8964	0.9176	+2.4%
Kiln	F1-macro	0.7394	0.7615	+3.0%
	Accuracy	0.8689	0.8777	+1.0%
Glaze	F1-macro	0.7338	0.7079	−3.5%
	Accuracy	0.8115	0.7940	−2.2%
Type	F1-macro	0.7463	0.7791	+4.4%
	Accuracy	0.8240	0.8352	+1.4%

The training histories from the Real+Syn-2500 configuration reveal distinct task-specific learning behaviors that illuminate why certain tasks benefit while others suffer

from synthetic augmentation. As shown in Table 4.21, the convergence patterns vary significantly across tasks. Dynasty and Type classifications tasks exhibit stable learning trajectories, reaching 90% of their optimal performance within 9–12 epochs despite the mixed data distribution. In contrast, Glaze classification exhibits significant instability (variance = 0.91), requiring 18 epochs to reach comparable performance levels. This variability is reflected in oscillating validation accuracy throughout training, suggesting that the model has difficulty reconciling authentic texture patterns with their synthetic approximations.

Table 4.21. Task-Specific Learning Dynamics with Syn-2500

Task	Epochs to 90%	Best Epoch	Val Loss Variance	Final Val Acc	Stability
Dynasty	9	46	0.42	0.938	Stable
Kiln	14	46	0.58	0.938	Moderate
Glaze	18	46	0.91	0.837	Unstable
Type	12	46	0.35	0.854	Very Stable

The most pronounced benefit of synthetic augmentation is observed in minority class performance, as shown in Figure 4.14 and quantified by per-class F1 scores derived from the confusion matrices. Table 4.22 presents the detailed breakdown of these differential impacts. These overall task-level differences are further reflected in minority–majority class behavior.

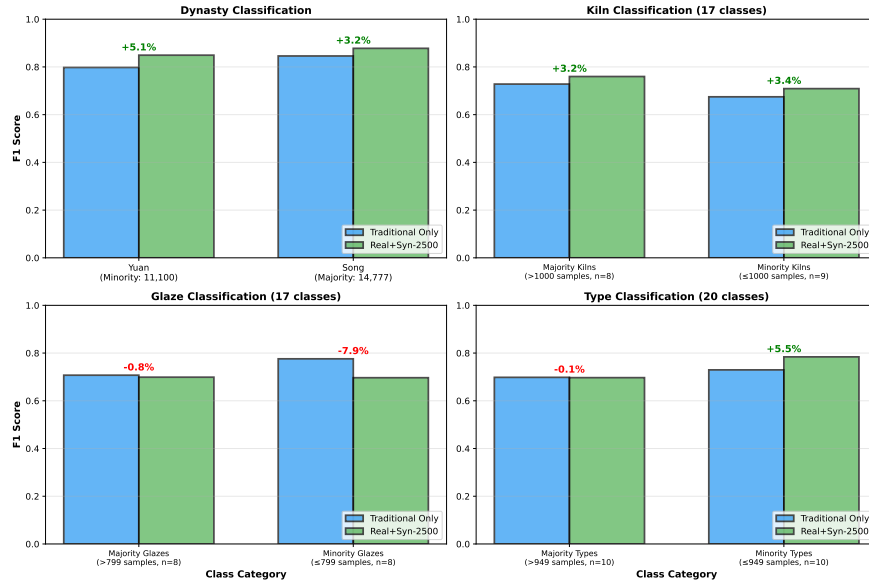


Figure 4.14. Minority vs. Majority Class Performance Gains with Synthetic Augmentation

Table 4.22. Minority vs. Majority Class F1 Improvements

Task	Class Type	Samples	Baseline F1	+Syn-2500 F1	Improvement
Dynasty	Yuan (minority)	11,100	0.794	0.844	+5.1%
	Song (majority)	14,777	0.845	0.876	+3.2%
Kiln	Majority (>1000)	8 classes	0.734	0.759	+3.2%
	Minority (≤ 1000)	9 classes	0.681	0.706	+3.4%
Glaze	Majority (>799)	8 classes	0.708	0.702	-0.8%
	Minority (≤ 799)	8 classes	0.773	0.703	-7.9%
Type	Majority (>949)	10 classes	0.703	0.702	-0.1%
	Minority (≤ 949)	10 classes	0.732	0.778	+5.5%

The results indicate task-dependent patterns in the impact of synthetic augmentation on class balance. In dynasty classification, the minority Yuan class benefits more substantially (+5.1%) than the majority Song class (+3.2%). Type classification exhibits an even greater disparity, with minority types gaining 5.5% while majority types remain essentially unchanged (-0.1%).

For kiln classification, improvements are similar for both majority and minority classes (~ 3.2 – 3.4%), suggesting that regional porcelain features benefit uniformly from synthetic augmentation regardless of sample size. In contrast, glaze classification shows an inverse pattern, with minority glazes experiencing a larger performance decrease (-7.9%) than majority glazes (-0.8%), indicating that the generation process poorly captures rare glaze textures.

The heterogeneous effects across tasks appear to be closely associated with their reliance on different visual feature types. Table 4.23 illustrates this relationship by mapping actual performance outcomes to the predominant visual features that each task relies on. The 4.4% improvement in type classification, the highest among all tasks, suggests that morphological diversity is effectively captured by the LoRA-adapted diffusion model. Tasks relying on global structural features consistently benefit from synthetic data’s capacity to generate diverse viewpoints while preserving overall proportions. In contrast, glaze classification’s dependence on subtle surface textures, which are often smoothed during the diffusion process, likely accounts for its consistent performance degradation across both majority and minority classes.

Table 4.23. Visual Feature Dependencies and Synthetic Impact

Visual Feature	Critical For	Actual Im- pact	Evidence from Data
Global shape/profile	Dynasty, Type	+3.9%, +4.4%	Highest F1 improvements
Surface texture	Glaze	-3.5%	Consistent degradation
Decorative patterns	Kiln	+3.0%	Uniform improvement
Proportions/symmetry	Type	+4.4%	Best overall performance
Regional markers	Kiln	+3.0%	Stable across classes

Part of these differences may stem from the structured prompt engineering used during generation, which was derived from established archaeological documentation methods described in Section 3.2.2. Each prompt followed a hierarchical template informed by traditional morphological authentication principles and museum cataloging standards:

Yuan dynasty, Jun kiln produced, Chinese porcelain, vase for display only,
no spout or handle, decorative vessel, with moon white glaze with pale blue
lighter than bluish green, thick opaque glaze

This structured approach incorporates five key archaeological descriptors. **Temporal attribution** (“Yuan dynasty”) was effectively captured in visual features, contributing to a +3.9% F1-macro improvement in dynasty classification; period markers such as vessel proportions and rim profiles, documented in Section 3.2.2, were successfully translated into visual characteristics. **Production site** (“Jun kiln produced”) was moderately captured, yielding a +3.0% gain in kiln classification; distinctive traits such as the Jun kiln’s milky opacity were partially reproduced, although confusion persisted for kilns with similar glazing techniques. **Functional classification** (“rice bowl for daily meals” vs. “vase for display only”) achieved the largest improvement (+4.4% in type classification); explicit functional descriptions aligned with morphological authentication principles, reflecting the traditional “form follows function” framework in scholarship. **Morphological descriptors** (“rounded shape, larger size”, “no spout or handle”) were strongly correlated with visual performance gains, consistent with the detailed form categories in Figure 3.3, specific shape variations (e.g., “angled waist”, “lobed sides”) had clear visual manifestations. Finally, **Glaze characterization** (“moon white glaze with pale blue lighter than bluish green, thick opaque glaze”) showed the lowest performance change (−3.5% in glaze classification), despite precise linguistic specification drawn from the glaze color categories in Figure 3.2, fine-grained texture and color gradations proved difficult for the generative model to reproduce.

The observed performance hierarchy (Type > Dynasty > Kiln > Glaze) aligns with the level of abstraction in the described features. Concrete morphological descriptors,

such as “phoenix-shaped vase” or “sugarcane-shaped washer” from the archaeological taxonomy, are consistently translated into reliable visual features. In contrast, more subjective qualities, including “milky opacity” or “crab claw pattern” used to define glaze characteristics, require texture-rendering capabilities that exceed the capacity of current diffusion models. This pattern underscores both the value and the limitations of archaeological domain knowledge in prompt engineering. While prompts incorporating specialized terminology, such as “purple rim and iron foot” (Guan kilns) or “gold wire and iron thread” (Ge kilns), accurately convey expert concepts, the generative process fails to reproduce these surface-level authentication markers that are critical for expert classification.

Prompt-guided generation effectively reduced several major confusions. For instance, the **TeaBowl**→**Dish** error dropped by 30 percentage points when prompts explicitly distinguished “tea ceremony vessel” from “serving dish”, while the **Yuan**→**Song** confusion decreased by 7 points with the inclusion of period-specific traits. In contrast, the **IvoryWhite**→**White** confusion showed little change despite detailed color descriptions, underscoring the model’s persistent limitations in texture rendering. These results indicate that prompt optimization alone cannot overcome texture synthesis constraints. Future directions include **hierarchical prompting** that prioritizes reliably translated features such as shape and function; **task-specific templates** tailored to classification requirements; and the **integration of materials science terminology** (e.g., “kaolin body,” “vitreous glaze”) to impose more precise generative constraints. Overall, while domain expertise enhances prompt design, successful augmentation requires recognizing the boundary between linguistically specifiable features and those that must be visually preserved.

Detailed analysis of the confusion matrices indicates that synthetic augmentation alters the distribution of classification errors in task-specific ways (Talbe 4.24). The largest improvement is observed in type classification, where TeaBowl→Dish confusion decreases by 30 percentage points, suggesting that synthetic data enhances shape discrimination through increased viewpoint diversity. In dynasty classification, Yuan→Song confusion decreases by 7 points, indicating consistent gains. In contrast, glaze classification shows bidirectional degradation, with White→IvoryWhite confusion increasing by 9 points, implying that synthetic generation may obscure subtle surface differences rather than resolve them.

Table 4.24. Key Confusion Pattern Changes with Syn-2500

Task	Confusion Pair	Baseline	+Syn-2500	Change	Interpretation
Dynasty	Yuan→Song	31%	24%	−7%	Better period discrimination
Kiln	LinC→Ding	17%	8%	−9%	Improved regional distinction
	Jian→Jian	67%	100%	+33%	Better rare kiln recognition
Glaze	IvoryW→White	35%	34%	−1%	Minimal improvement
	White→IvoryW	16%	25%	+9%	Increased reverse confusion
Type	TeaB→Dish	70%	40%	−30%	Major shape clarification
	Pot→Vase	10%	19%	+9%	New confusion introduced

Beyond confusion pairs, Top-5 accuracy offers an additional perspective on whether synthetic augmentation alters prediction confidence rather than overall class ranking. Analysis of the Top-5 accuracy metrics illustrates the nature of the improvements achieved through synthetic augmentation (Table 4.25). The minimal changes in Top-5 accuracy (maximum 0.63%) indicate that synthetic augmentation primarily influences classification confidence rather than the model’s capacity to identify plausible alternatives. This pattern implies that synthetic data either provides clear discriminative features that enhance Top-1 accuracy or introduces ambiguity that reduces it, while exerting minimal effect on the relative ranking of alternative predictions. In glaze classification, the slight improvement in Top-5 accuracy despite a decline in Top-1 performance suggests that training with synthetic data reduces prediction confidence, leading to a more even distribution of probability mass across similar classes.

Table 4.25. Top-5 Accuracy Analysis

Task	Baseline Top-5	+Syn-2500 Top-5	Change	Interpretation
Dynasty	N/A (binary)	N/A	-	-
Kiln	0.9863	0.9875	+0.12%	Marginal improvement
Glaze	0.9800	0.9863	+0.63%	Small benefit for alternatives
Type	0.9738	0.9738	0%	No change in near-misses

The divergent task-specific effects revealed by the confusion matrix analysis raise questions about optimal architecture design for multi-task learning with synthetic augmentation. In the current uniform configuration, where all tasks process identical augmented datasets, type classification achieves a 4.4% gain at the cost of a 3.5% decline in glaze classification performance. Based on these findings, three architectural modifications are proposed: (1) **Task-specific data pipelines**, where dynasty, kiln, and type branches leverage synthetic augmentation while the glaze branch processes only authentic images, thereby preserving the 3.9–4.4% gains while avoiding the 3.5% loss; (2) **Adaptive loss weighting**, dynamically adjusting task weights to reduce glaze-task influence during synthetic batches; and (3) **Feature-aware generation**,

developing synthesis models optimized for distinct visual features, since general-purpose synthesis is suboptimal for texture-dependent tasks.

Taken together, these observations converge on a consistent conclusion. The analysis confirms that synthetic augmentation is beneficial only when generative fidelity aligns with task requirements. Morphological and stylistic features (Dynasty +3.9%, Type +4.4%, Kiln +3.0%) consistently improved, reflecting the model’s ability to capture shape and period markers. In contrast, glaze classification declined (−3.5%), with minority glazes suffering the largest drop (−7.9%), as subtle surface textures were poorly rendered. Error reductions such as TeaBowl→Dish (−30 pp) highlight augmentation’s value for morphology, while increased White→IvoryWhite confusion illustrates its limits for texture. These results demonstrate that synthetic augmentation is effective for structure-based tasks but counterproductive for texture-dependent classification, underscoring the need for selective rather than universal application.

4.6 Discussion and Conclusion

4.6.1 Summary of Key Findings

This chapter presented a comprehensive investigation into the integration of LoRA-enhanced synthetic data generation with traditional augmentation strategies for archaeological porcelain classification. Comparing a baseline of 25,877 traditionally augmented images against configurations with 570 (2.2%) and 2,500 (8.8%) synthetic images, we established both the potential and limitations of synthetic augmentation in this domain. Overall, the Real+Syn-2500 configuration achieved a marginal 1.5% gain in F1-macro (0.767→0.782), yet the effect was highly task-dependent: dynasty (+3.9%), type (+4.4%), and kiln (+3.0%) benefited consistently, while glaze declined by 3.5%. This divergence highlights that synthetic augmentation is effective primarily for structure- and morphology-based tasks but counterproductive for texture-dependent classification. A clear threshold effect was observed: the Syn-570 configuration produced negligible or even negative impact, whereas at least 8–9% synthetic data was required for measurable improvements, indicating that synthetic samples must form a sufficient proportion of training data to reshape the learned distribution. Per-class analysis further revealed that minority categories gained most, with Yuan dynasty (+5.1%) and minority vessel types (+5.5%) improving substantially, while minority glazes suffered a sharp decline (−7.9%), underscoring the difficulty of synthesizing rare texture patterns. Rigorous quality control was essential, with 8.8% of generated images rejected for historical or semantic inconsistencies, yielding a 91.2% acceptance rate but necessitating 10%

overproduction. Finally, structured prompt engineering grounded in archaeological documentation proved effective for morphological descriptors (e.g., “rounded shape”), which reliably enhanced type classification, but failed to capture fine-grained glaze characteristics (e.g., “milky opacity”), confirming that linguistic specification cannot overcome fundamental limitations in texture synthesis. Together, these findings demonstrate that synthetic augmentation is a selective rather than universal tool: when aligned with task-relevant visual features, it improves representation and classification, but its benefits are constrained by the representational boundaries of current generative models.

4.6.2 Theoretical and Practical Implications

This work advances the understanding of synthetic-real distribution relationships in domain-specific computer vision tasks. The observed task-dependent performance variations provide empirical evidence that the effectiveness of synthetic data depends not on dataset size or class balance alone, but on the alignment between what generative models can faithfully reproduce and the discriminative features required by each task. The identified threshold at 8–9% synthetic proportion suggests a phase transition in which synthetic patterns attain sufficient representation to influence feature learning, contributing to theories of mixed training dynamics in neural networks.

The differential impact across visual features, benefiting shape-based tasks while hindering texture-based tasks, supports the hypothesis that current diffusion models have an inherent smoothing bias, preserving global structure but degrading fine-grained surface details. This finding implies that synthetic augmentation strategies should be tailored to the specific visual features critical for each application domain.

This work provides empirical evidence for the limits of prompt-guided generation in specialized domains, showing that translation success decreases as feature abstractness increases. Despite the use of archaeologically grounded terminology from museum catalogs and scholarly documentation, such as “gold wire and iron thread” for Ge kilns or “fish scale crackles” for Ru kilns, performance varied by feature type: concrete morphological descriptors achieved high fidelity (+4.4% in type classification), period markers showed moderate success (+3.9% in dynasty classification), and surface texture descriptions consistently failed (−3.5% in glaze classification). These findings contribute to understanding the semantic gap between expert linguistic descriptions and their visual realization in generative models.

The observed hierarchy of translation success, morphology > period style > regional characteristics > surface texture, indicates that generative models are more effective

at reproducing structural features than material properties. This finding clarifies the practical boundaries of synthetic data use in archaeological porcelain research and suggests that, while domain expertise is essential for prompt engineering, it cannot compensate for architectural limitations in texture synthesis.

These theoretical insights directly inform practical guidelines for the deployment of synthetic augmentation in archaeological porcelain research. Based on the empirical findings, we recommend a decision framework for the use of synthetic augmentation in archaeological porcelain classification. Synthetic augmentation should be applied when classification tasks rely primarily on morphological or structural features, when minority classes contain fewer than 100 samples after traditional augmentation, and when computational resources permit 3-4 hours of LoRA training. In such cases, even marginal improvements of 3-4% can justify the additional computational cost. Conversely, synthetic augmentation should be avoided when texture fidelity is critical (e.g., glaze classification), when datasets already exceed 100 samples per class, when available computational resources are restricted to under 1 hour, or when quality control for generated data cannot be ensured. Effective prompt engineering should emphasize concrete morphological descriptors (e.g., "rounded shape", "larger size") and functional categories (e.g., "rice bowl for daily meals" vs. "vase for display only"), while using established archaeological terminology for dynasty and kiln attribution. However, detailed glaze descriptions such as "thick opaque glaze" cannot compensate for current generative model limitations in texture synthesis. Finally, the optimal configuration targets an 8-9% synthetic proportion, adopts task-specific data pipelines in multi-task settings, and prioritizes augmentation of minority classes with structural rather than textural features.

4.6.3 Limitations and Future Work

The primary limitation of the current approach is the unresolved issue of texture degradation, with glaze classification consistently exhibiting performance losses. The inherent smoothing in the diffusion model's denoising process conflicts with the preservation of fine surface details. Despite sophisticated prompt engineering grounded in archaeological expertise, the language-to-visual translation remains constrained by the model's texture synthesis capabilities. Even precise terminology, such as "purple rim and iron foot" (Guan kilns) or "rabbit hair striation" (Jian kilns), fails to produce authentic surface characteristics.

Several promising avenues for further investigation emerge from these findings. First, task-adaptive generation could be developed by training separate LoRA adaptations

optimized for distinct visual features, such as one specialized in shape and structure and another focused on texture and surface, to address the current trade-off between tasks. Second, multi-modal prompt enhancement may prove valuable by combining textual prompts with texture patches or technical specifications (e.g., surface roughness measurements or spectral reflectance data), thereby bridging the gap between linguistic descriptions and visual realizations. Third, the development of archaeological language models trained on museum catalogs, excavation reports, and scholarly literature could enable more accurate interpretation of domain-specific terminology, capturing subtle associations between descriptors such as *celadon* and their visual traits. Fourth, texture-preserving architectures like VQ-VAE or dedicated texture synthesis networks should be explored to maintain fine-grained surface details that are critical for glaze classification. Finally, dynamic proportion optimization represents another promising direction, where adaptive algorithms determine the optimal proportion of synthetic data for each task based on validation performance, potentially leveraging reinforcement learning to balance multi-objective performance.

As generative models advance, particularly in texture synthesis and fine-detail preservation, the role of synthetic data in archaeological porcelain applications is likely to expand. Such expansion, however, should be guided by task-specific evaluation, domain expertise, and a clear understanding of the translation gap between linguistic description and visual realization. The contribution of this work lies in two areas: first, demonstrating the feasibility of integrating synthetic augmentation in archaeological porcelain classification; and second, establishing quantitative frameworks to determine both when synthetic augmentation delivers value in specialized domains and when linguistic specification reaches its limits in guiding visual generation.

Beyond technical limitations, the use of synthetic data in archaeological contexts raises fundamental epistemological questions. While our approach demonstrates that visually similar synthetic images can improve CNN performance, visual similarity does not equate to archaeological validity. Synthetic images, regardless of their photorealistic quality, are representations without material reality, historical provenance, or manufacturing context. This raises two principal risks: first, the potential hallucination of historically inaccurate features, such as motifs, glaze colours, or stylistic combinations that are visually plausible but unattested in historical production; and second, the risk that models trained on synthetic data primarily learn surface-level visual regularities rather than features grounded in authentic production techniques, materials science, or cultural practices.

It is therefore essential to distinguish between visual classification, the computa-

tional task addressed in this work, and archaeological attribution, which is a scholarly process integrating materials analysis (e.g., clay composition, firing temperature, glaze chemistry), provenance documentation, stylistic development, and expert connoisseurship. The former operates on statistical regularities in pixel distributions, whereas the latter relies on multi-modal evidence that cannot be captured by computer vision alone. Accordingly, our method is intended for preliminary sorting, educational applications, or exploratory dataset augmentation; however, it cannot substitute for the evidential standards required for museum accessioning, conservation practices, or scholarly authentication.

To mitigate these epistemological risks, future work should incorporate expert validation, whereby porcelain specialists assess the historical plausibility of synthetic samples and transparent reporting practices that explicitly label synthetic augmentation and quantify associated uncertainties. As generative models continue to advance, the archaeological community must develop shared standards for their responsible use, ensuring that computational tools enhance rather than obscure the material, historical, and cultural complexity of archaeological artifacts.

Chapter 5

Conclusion

This thesis investigated the application of machine learning techniques for automated classification of Chinese porcelains, addressing the critical challenge of limited annotated data in archaeological porcelain studies. Through a systematic exploration of transfer learning, data augmentation via synthetic image generation, and multi-task learning frameworks, this research demonstrated that deep learning models can achieve robust performance even with constrained datasets. The integration of MobileNetV3 architecture with progressive fine-tuning strategies yielded classification accuracies exceeding 85% across four distinct tasks: dynasty attribution, kiln identification, glaze classification, and vessel type recognition. The successful deployment of Stable Diffusion for synthetic data generation, producing 1,000 photorealistic porcelain images, proved particularly effective in addressing class imbalance and improving model generalization, establishing a novel approach for data augmentation in cultural heritage applications. This chapter summarises the contributions of this thesis, highlighting key findings for each research question we presented in Chapter 1.

RQ-1: *How can machine learning methods be effectively applied to the identification of Chinese archaeological porcelain, given the current limitations of available data?*

RQ1.1: *What is the state of the art in machine-learning-based identification of archaeological ceramics, and what methodological or data-related gaps remain?*

RQ1.2: *What types of datasets, image-processing techniques, and learning algorithms have proven most effective in previous studies, and how do they inform the design of a robust porcelain-classification framework?*

RQ1.3: *What are the main ML algorithms used in the identification of archaeological ceramics?*

These research questions are investigated through a systematic literature review in Chapter 2. This review examines the trends of machine learning in the identification of archaeological ceramics. Specifically, it focuses on the application of ML and DL algorithms between 2013 and 2023, on several types of ceramic (pottery and porcelain). It investigates several aspects such as the available datasets, the image processing

methods, and ML or DL algorithms. In terms of the application of ML algorithms to the image classification of ceramic images, papers from 2013 to 2020 tend to use traditional ML algorithms such as SVM and KNN. However, there is a general shift towards the use of DL based on CNN models for classification in papers from 2020 to 2023. In fact, based on the conclusions drawn from these papers, DL algorithms have made significant breakthroughs in many tasks due to their powerful automatic feature extraction capabilities. Among CNN models, VGG16, MobileNet and ResNet have achieved good performance on archaeological ceramic images and higher accuracy compared to traditional ML algorithms such as SVM and KNN. This indicates that DL algorithms such as CNN models will be the main trend in the future for archaeological ceramic classification applications. Various studies demonstrated the technical potential of DL for the identification of archaeological ceramics. The manual classification of a large number of pottery sherds is a time-consuming activity and empirical identification methods of porcelains often lack objective constraints. The image classification based on DL provides support for the classification of archaeological ceramics. This review and its conclusions should be seen as a first step in the analysis of machine learning for ceramic identification in the sector.

Overall, the use of machine learning algorithms in archaeological ceramics classification complements traditional methods by offering efficiency, accuracy, and new avenues for data analysis and interpretation. By combining the strengths of both approaches, researchers can gain deeper insights into past civilizations and cultural practices. The efficiency of these algorithms can speed up the classification process allowing archaeologists to analyze larger datasets and automate repetitive tasks, such as sorting and categorizing artifacts based on specific attributes. This automation frees up archaeologists' time to focus on higher-level analysis and interpretation. Moreover, machine learning algorithms can uncover hidden relationships or correlations within ceramic datasets that may not be apparent through traditional methods. By analyzing these insights, archaeologists can gain a deeper understanding of cultural practices, trade networks, and historical contexts.

RQ-2: *To what extent can deep learning models (especially CNN-based multi-task architectures) accurately classify key attributes of Chinese porcelain under real data constraints?*

RQ2.1: *What performance can be achieved when simultaneously classifying dynasty, kiln provenance, glaze, and vessel form?*

RQ2.2: *How does transfer learning influence classification performance under limited, imbalanced archaeological datasets?*

RQ2.3: *How can explainable AI methods help interpret model decisions and relate computational features to archaeologically meaningful attributes?*

These research questions are explored through Chapter 3. This study demonstrates that multi-task learning with transfer learning provides an effective framework for automated porcelain classification, achieving performance levels that approach practical deployment requirements while revealing fundamental challenges in fine-grained cultural heritage analysis. The comprehensive evaluation across 801 test images representing 141 unique porcelain combinations establishes both the capabilities and limitations of current deep learning approaches for archaeological artifact authentication. The consistent performance of MobileNetV3 across all four tasks (standard deviation of only 0.049 in F1 scores) demonstrates balanced multi-task optimization without sacrificing individual task performance. This stability contrasts with the higher variance observed in other architectures, particularly EfficientNet-B2, suggesting that lightweight models with carefully designed architectures may be better suited for multi-objective optimization than simply scaling up model capacity. The multi-task framework demonstrates clear advantages over single-task approaches, with shared representations enabling effective learning from the augmented training set (25,877 samples providing supervision signals for four different objectives simultaneously). The high cross-task attention correlation (0.98–1.00) confirms successful feature sharing, while task-specific performance variations indicate appropriate specialization within the shared backbone. The Grad-CAM analysis provides crucial validation that the model learns archaeologically meaningful features rather than dataset artifacts. The attention patterns focusing on decorations for *Dynasty* classification, glaze surfaces for material identification, and vessel profiles for morphological categorization mirror the diagnostic features used by human experts. This alignment between learned and expert features strengthens confidence in the model’s decision-making process and suggests potential for explainable AI applications in cultural heritage contexts.

This comprehensive evaluation establishes multi-task learning with lightweight transfer learning architectures as a viable approach for automated porcelain classification, achieving performance levels suitable for practical deployment while maintaining interpretability crucial for cultural heritage applications. The success of **MobileNetV3**, balancing high accuracy, parameter efficiency, and fast training, demonstrates that carefully designed efficient architectures can match or exceed the performance of larger models while offering superior deployment characteristics.

RQ-3: *Can synthetic image generation mitigate data scarcity and imbalance in archaeological porcelain datasets without compromising classification reliability?*

RQ3.1: *How do the quality and diversity of generated images affect CNN training?*

RQ3.2: *Can synthetic images effectively balance underrepresented classes and improve performance on rare kiln, glaze, or form categories?*

RQ3.3: *What is the optimal combination of real and synthetic data to maximise model stability and generalisation?*

These research questions are discussed in Chapter 4. This study presents a comprehensive investigation into the integration of LoRA-enhanced synthetic data generation with traditional augmentation strategies for archaeological porcelain classification. Comparing a baseline of 25,877 traditionally augmented images against configurations with 570 (2.2%) and 2,500 (8.8%) synthetic images, we established both the potential and limitations of synthetic augmentation in this domain. Overall, the Real+Syn-2500 configuration achieved a marginal 1.5% gain in F1-macro (0.767→0.782), yet the effect was highly task-dependent: dynasty (+3.9%), type (+4.4%), and kiln (+3.0%) benefited consistently, while glaze declined by 3.5%. This divergence highlights that synthetic augmentation is effective primarily for structure- and morphology-based tasks but counterproductive for texture-dependent classification. A clear threshold effect was observed: the Syn-570 configuration produced negligible or even negative impact, whereas at least 8–9% synthetic data was required for measurable improvements, indicating that synthetic samples must form a sufficient proportion of training data to reshape the learned distribution. Per-class analysis further revealed that minority categories gained most, with Yuan dynasty (+5.1%) and minority vessel types (+5.5%) improving substantially, while minority glazes suffered a sharp decline (−7.9%), underscoring the difficulty of synthesizing rare texture patterns. Rigorous quality control was essential, with 8.8% of generated images rejected for historical or semantic inconsistencies, yielding a 91.2% acceptance rate but necessitating 10% overproduction. Finally, structured prompt engineering grounded in archaeological documentation proved effective for morphological descriptors (e.g., “rounded shape”), which reliably enhanced type classification, but failed to capture fine-grained glaze characteristics (e.g., “milky opacity”), confirming that linguistic specification cannot overcome fundamental limitations in texture synthesis. Together, these findings demonstrate that synthetic augmentation is a selective rather than universal tool: when aligned with task-relevant visual features, it improves representation and classification, but its benefits are constrained by the representational boundaries of current generative models.

For future research directions, this research opens several promising avenues for future investigation in assisting automated porcelain authentication and broader applications in cultural heritage preservation. The immediate extension involves expanding the dataset

to encompass porcelains from underrepresented dynasties (Tang and Five Dynasties) and non-Chinese porcelains, enabling cross-cultural comparative studies and testing the transferability of learned features across distinct porcelain traditions. The synthetic data generation pipeline could be enhanced through fine-tuning diffusion models on larger archaeological datasets, potentially incorporating 3D reconstruction techniques to generate multi-view representations that capture morphological features invisible in single photographs, such as foot ring profiles and wall thickness variations critical for authentication. The broader implications extend to other archaeological materials, including bronze vessels, jade artifacts, and ancient textiles, where similar data scarcity challenges exist. The synthetic data generation methodology could revolutionize training data creation across cultural heritage domains, while the interpretability techniques developed here could establish standards for explainable AI in archaeological applications, building trust among domain experts traditionally skeptical of automated approaches. Ultimately, this research validates that synthetic data generation and multi-task learning can partially mitigate data scarcity challenges in archaeological AI applications, providing a reproducible framework for similar cultural heritage challenges while ensuring model decisions remain interpretable and aligned with established scholarly practices through AI-assisted rather than AI-replaced workflows.

I

Appendix

Chapter A

Appendix A

Table A.1. Details of the reviewed studies

ID	Paper Title	Published Year
ID1 [55]	Technological innovation in the recognition process of Yaozhou Kiln ware patterns based on image classification	2023
ID2 [69]	Classification of unexposed potsherd cavities by using deep learning	2023
ID3 [100]	Identification of porcelain ewers in Tang, Song, and Yuan dynasties by digital shape characterization	2023
ID4 [118]	Comparative Study on Image Classification Models of Ancient Ceramic Types	2023
ID5 [26]	Ceramic Texture Extraction Method Based on Convolution Neural Network and Image Processing	2023
ID6 [30]	Typical Ceramic Pattern Extraction Based on Convolutional Neural Network	2023
ID7 [76]	Using Image Feature Extraction to Identification of Ancient Ceramics Based on Partial Differential Equation	2022
ID8 [56]	Automatic Classification of Blue and White Porcelain Sherds Based on Data Augmentation and Feature Fusion	2022
ID9 [10]	A deep variational convolutional Autoencoder for unsupervised features extraction of ceramic profiles. A case study from central Italy	2022
ID10 [48]	Ceramic Decoration Extraction Method Based on Computer Vision and Image Processing	2022
ID11 [19]	Intelligent Dating of Chinese Ancient Ceramics Based on Convolutional Neural Network	2022
ID12 [47]	Using Deep Learning for the Image Recognition of Motifs on the Center of Sukhothai Ceramics	2022
ID13 [116]	Ceramic Type Recognition Algorithm Based on Ontology Modeling and Transfer Learning	2022
ID14 [79]	Applications of deep learning to decorated ceramic typology and classification: A case study using Tusayan White Ware from Northeast Arizona	2021
ID15 [117]	Smart Culture Lens: An Application That Analyzes the Visual Elements of Ceramics	2021
ID16 [61]	Identification Method of Ancient Ceramics Revision	2021
ID17 [33]	The Application of Computer Vision in the Detection of Ancient Ceramic Art	2021
ID18 [60]	Ceramic Fabric Classification of Petrographic Thin Sections with Deep Learning	2021
ID19 [17]	Applications of learning methods to imaging issues in archaeology, regarding ancient ceramics manufacturing	2021
ID20 [119]	Characteristics of Banana Leaf Pattern for Ceramics in Yuan, Ming and Qing Dynasties	2020
ID21 [73]	Research on ancient ceramics identification by artificial intelligence	2019
ID22 [126]	A Framework for Design Identification on Heritage Object	2019
ID23 [34]	Automated classification of archaeological ceramic materials by means of texture measures	2018
ID24 [8]	Machine learning identification and classification of historic ceramics	2018
ID25 [44]	Using supervised machine learning to classify ceramic fabrics	2018
ID26 [63]	Classification Archaeological Fragments into Groups	2017
ID27 [127]	Porcelain image classification based on semi-supervised mean shift clustering	2017
ID28 [110]	Machine Vision Based Classification and Identification for Non-destructive Authentication of Ancient Ceramic	2017
ID29 [85]	Using Both HSV Color and Texture Features to Classify Archaeological Fragments	2015
ID30 [105]	Automatic pattern recognition on archaeological ceramic by 2D and 3D image analysis: A feasibility study	2015
ID31 [80]	Towards the automatic classification of pottery sherds: two complementary approaches	2015
ID32 [3]	A Semantic Analysis Approach to Thin-Shell Ceramic Fragment Classification	2013
ID33 [65]	Automatic classification of archaeological pottery sherds	2013

Table A.2. Dataset details of the reviewed studies

ID	Ceramic Type	Dataset
ID1 [55]	Porcelain	This study analyzed Yaozhou kiln ceramic from China. The production period is not provided.
ID2 [69]	Pottery	This study analyzed archaeological pottery fragments from the Late Jomon period (ca.4000 to 3200 aBP) in Japan. The proposed dataset is composed of 1,036 X-ray images from seven classes.
ID3 [100]	Porcelain	This study selected images of porcelain ewers from the Tang (ca.1332 to 1043 aBP), Song (ca.990 to 671 aBP), and Yuan (ca.679 to 582 aBP) dynasties in China. There were 85 images from the Tang Dynasty, 85 from the Song Dynasty, and 62 from the Yuan Dynasty. Images were photographed at museums in China.
ID4 [118]	Porcelain	The study used the porcelain image dataset provided by Baidu PaddlePaddle to collect archaeological porcelain images in the datasets, including 19 different types of vases in China. The production period is not provided.
ID5 [26]	Ceramic	Not provided
ID6 [30]	Porcelain	The datasets are mainly derived from Chinese museums' photo scans and websites. Four typical porcelain pattern datasets of dragon, fish, lotus, and peony are established. The production period is not provided.
ID7 [76]	Porcelain	The study extracted and validated 5,834 images of archaeological ceramics from China, including 272 types of kilns, Yaozhou kiln, and Yue kiln with manual labelling. The production period is not provided.
ID8 [56]	Porcelain	This paper employed blue and white porcelain shards by incomplete patterns, a total of 373 ceramic image samples of 14 types in the Yuan (ca.679 to 582 aBP), Ming (ca.582 to 306 aBP) and Qing (ca.306 to 38 aBP) dynasties from China.
ID9 [10]	Pottery	This study focused on Lazio and Etruria pottery, a dataset of 4992 complete vessel profiles from 117 sampled contexts between the Late Bronze Age and the Orientalizing period (ca.3100 to 2530 aBP) in Italy.
ID10 [48]	Ceramic	Not provided
ID11 [19]	Porcelain	This paper takes Mei vase and Yuhuchun vase in Ming (ca.582 to 306 aBP) and Qing (ca.306 to 38 aBP) dynasties as the research object, all samples are from the Beijing Museum and Taipei Museum website. A total of 248 available samples were obtained.
ID12 [47]	Ceramic	The images in the dataset were obtained by photographing the ceramics of Sukhothai kilns in many private museums. A total of 557 images that featured seven types of patterns were collected from Thailand. Not provided the production period of the ceramics.
ID13 [116]	Porcelain	The dataset of the paper was collected from the official website of the Palace Museum in China. A total of 2750 pictures of the most representative ceramics, color glaze, under-glaze color, blue and white porcelain, purple sand, and doucai from Song (ca.990 to 671 aBP), Yuan (ca.679 to 582 aBP), Ming (ca.582 to 306 aBP) and Qing (ca.306 to 38 aBP) dynasties.
ID14 [79]	Pottery	A set of 3064 Tusayan White Ware(TWW) sherd photographs for the dataset from regional museum collections in the United States of America (ca.1125 to 650 aBP), providing a rich source of imagery due to the contrast between the paint and background of the sherds.
ID15 [117]	Ceramic	The dataset required for this study was created using two approaches. First, A total of 5,684 images of 5,211 ceramics were collected by the National Museum. Second, 13,926 Pictures of 2,135 ceramics were selected at 19 institutions in the Republic of Korea. Not provided the production period of the ceramics.
ID16 [61]	Porcelain	The self-built porcelain set includes seven types, 5624 pieces of porcelain bottles which had been crawled through Baidu in China. Not provided the production period of the porcelains.
ID17 [33]	Ceramic	Not provided
ID18 [60]	Pottery	The selected dataset used was primarily composed of 66 pottery sherds images of petrographic thin sections under cross-polarized light from the Cocal-period archaeological site of Peten from Honduras (ca.950 to 425 aBP). In total, 1069 images were used to train each model.
ID19 [17]	Pottery	The dataset selected in this study was composed of sixteen sherds: four coiling technology (CoT) sherds and twelve spiralled patchwork technology (SPT) sherds. The dataset originates from the early Neolithic period (ca.7850 to 7450 aBP) in southeastern France.
ID20 [119]	Porcelain	A total of 230 pieces of ancient porcelains containing banana leaf patterns from Yuan (ca.679 to 582 aBP), Ming (ca.582 to 306 aBP) and Qing (ca.306 to 38 aBP) dynasties, collected from museums in China.
ID21 [73]	Porcelain	Analyzed the porcelain washer from Ming dynasty (ca.582 to 306 aBP) in China.
ID22 [126]	Pottery	The dataset consists of curvilinear paddle impressions on pottery sherds and comes from North America, specifically during the period (ca.1600 to 1300 aBP)
ID23 [34]	Pottery	The dataset discussed in the paper involves archaeological pottery materials specifically sourced from two different types of Egyptian clays: Nile clay(ca.2250 to 2450 aBP) and Marl clay (ca.3250 to 3450 aBP). A total of the 69 samples were imaged under the optical microscope as thin sections using cross-polarized light.
ID24 [8]	Ceramic	The dataset includes historic ceramics recovered from archaeological sites in New Zealand. Not provided the production period of the ceramics.
ID25 [44]	Pottery	The dataset used in the study consists of pottery samples from Zulu communities in South Africa. The potteries are contemporary, made by modern potters who employ traditional methods that are reflective of historical practices.
ID26 [63]	Ceramic	This study selected eight classes of the Independence National Historical Park dataset(ca.175 to 125 aBP), a collection of ceramic fragments in the United States of America.
ID27 [127]	Porcelain	Dataset images consist of 288 porcelain pieces, including 13 different porcelains from China. Not provided the production period of the porcelains.
ID28 [110]	Porcelain	A total of 50 porcelain jar images were selected from Forbidden City Museum of China, including five ages of Neolithic (ca.12000 to 4000 aBP), Tang (ca.1332 to 1043 aBP), Song (ca.990 to 671 aBP), Yuan (ca.679 to 582 aBP) and Ming (ca.582 to 306 aBP).
ID29 [85]	Ceramic	This study selected six images of the Independence National Historical Park dataset (ca.175 to 125 aBP), a collection of ceramic fragments in the United States of America.
ID30 [105]	Pottery	The dataset used includes 935 shard imprints manually obtained by archaeologists (ca.1524 to 924 aBP) from France.
ID31 [80]	Pottery	A set of 203 pottery sherds collected from Greece. Not provided the production period of the potteries.
ID32 [3]	Ceramic	This paper selected ceramic images of the Independence National Historical Park dataset (ca.175 to 125 aBP), a collection of thin shell ceramic fragments in the United States of America.
ID33 [65]	Pottery	A set of 110 pottery sherd images including six different classes. The sherds were selected and photographed from Byzantine Antiquities in Greece (ca.1330 to 330 aBP).

Bibliography

- [1] Abadi, M., Khoudeir, M., & Marchand, S. 2012, in Image and Signal Processing: 5th International Conference, ICISP 2012, Agadir, Morocco, June 28-30, 2012. Proceedings 5, Springer, 333–342
- [2] Abdalla, H. B. 2022, Journal of Big Data, 9, 1
- [3] Aryafar, K., & Jeppson, P. L. 2013, International journal of heritage in the digital era, 2, 291
- [4] Baek, S., Lee, J.-H., Ahn, S., & Choi, J. 2025, ETRI Journal
- [5] Bao, Q., Zhao, J., Liu, Z., & Liang, N. 2025, Electronics, 14, 725
- [6] Bezur, A., & Casadio, F. 2012, Handheld XRF for art and archaeology, 249
- [7] Bi, H., Chen, Y., Chen, C., & Shu, L. 2025, Sensors (Basel, Switzerland), 25, 5558
- [8] Bickler, S. H. 2018, Archaeology in New Zealand, 20, 20
- [9] Bleichmar, D., & Mancall, P. C. 2011
- [10] Cardarelli, L. 2022, Journal of Archaeological Science, 144, 105640
- [11] Chen, Y., Zhang, A., Gao, F., & Guo, J. 2025, npj Heritage Science, 13, 205
- [12] Chong, M. J., & Forsyth, D. 2020, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 6070–6079
- [13] Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. 2019, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 9268–9277
- [14] Cumbajin, E., Rodrigues, N., Costa, P., et al. 2023, Journal of Imaging, 9, doi: [10.3390/jimaging9100193](https://doi.org/10.3390/jimaging9100193)
- [15] Deng, J., Dong, W., Socher, R., et al. 2009, in 2009 IEEE conference on computer vision and pattern recognition, Ieee, 248–255
- [16] Di Angelo, L., Di Stefano, P., & Guardiani, E. 2022, Journal of Cultural Heritage, 56, 10

- [17] Dia, K., Coli, V. L., Leblond, J., et al. 2021, 109
- [18] Elgammal, A., Liu, B., Kim, D., Elhoseiny, M., & Mazzone, M. 2018in
- [19] Feng, J., Zhou, Q., Zhang, R., Wang, Y., & LUO, H. 2022, Journal of Ceramics, 43, 145
- [20] Feng, X. 1982, History of Chinese Ceramics (Artifacts Publisher). <https://books.google.it/books?id=UOCGAAAAIAAJ>
- [21] Finlay, R. 2010, The pilgrim art: cultures of porcelain in world history, Vol. 11 (Univ of California Press)
- [22] Frid-Adar, M., Diamant, I., Klang, E., et al. 2018, Neurocomputing, 321, 321
- [23] Gai, K., Qiu, M., Zhao, H., & Sun, X. 2017, IEEE Transactions on Sustainable Computing, 3, 60
- [24] Ganin, Y., Ustinova, E., Ajakan, H., et al. 2016, Journal of machine learning research, 17, 1
- [25] Ganiyeva, F. 2023, Vakanüvis-Uluslararası Tarih Araştırmalar Dergisi, 8, 1513
- [26] Gao, Y. 2023, in 2023 IEEE 4th Annual Flagship India Council International Subsections Conference (INDISCON), IEEE, 1–6
- [27] Gîrbacia, F. 2024, Electronics, 13, 3738
- [28] Gong, T., Lee, T., Stephenson, C., et al. 2019, IEEE Access, 7, 141627
- [29] Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. 2020, Communications of the ACM, 63, 139
- [30] Han, L., Wang, Y., & Wang, X. 2023, in 2023 International Conference on Culture-Oriented Science and Technology (CoST), IEEE, 42–46
- [31] Haralick, R. M., & Shapiro, L. G. 1985, Computer vision, graphics, and image processing, 29, 100
- [32] He, K., Zhang, X., Ren, S., & Sun, J. 2016, in Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778
- [33] He, Y. 2021, Revista Argentina de Clinica Psicologica, 30, 325

- [34] Hein, I., Rojas-Domínguez, A., Ornelas, M., D’Ercole, G., & Peloschek, L. 2018, *Journal of Archaeological Science: Reports*, 21, 921
- [35] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. 2017, *Advances in neural information processing systems*, 30
- [36] Ho, J., Jain, A., & Abbeel, P. 2020, *Advances in neural information processing systems*, 33, 6840
- [37] Howard, A., Sandler, M., Chu, G., et al. 2019, in *Proceedings of the IEEE/CVF international conference on computer vision*, 1314–1324
- [38] Hu, E. J., Shen, Y., Wallis, P., et al. 2022, *ICLR*, 1, 3
- [39] Huang, T. S., Schreiber, W. F., & Tretiak, O. J. 1971, *Proceedings of the IEEE*, 59, 1586
- [40] Ioffe, S., & Szegedy, C. 2015, in *International conference on machine learning*, pmlr, 448–456
- [41] Kilis, N., Martinopoulou, E., Terzoglou, G., et al. 2025, *Digital Applications in Archaeology and Cultural Heritage*, e00451
- [42] Koh, J. Y., Fried, D., & Salakhutdinov, R. R. 2024, *Advances in Neural Information Processing Systems*, 36
- [43] Kornblith, S., Shlens, J., & Le, Q. V. 2019, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2661–2671
- [44] Koval, O. 2018
- [45] Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, *Advances in neural information processing systems*, 25
- [46] Kuang, Z., Zhang, J., Li, Y., & Fukuda, T. 2025, *npj Heritage Science*, 13, 256
- [47] Kuntitan, P., & Chaowalit, O. 2022, *Current Applied Science and Technology*, 10
- [48] Li, L. 2022, 388
- [49] Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. 2021, *IEEE transactions on neural networks and learning systems*, 33, 6999
- [50] Lin, J. 2025, *International Journal of Innovative Computing and Applications*, 15, 192

- [51] Ling, Z., Delnevo, G., Salomoni, P., & Mirri, S. 2025, arXiv preprint arXiv:2503.14231
- [52] Ling, Z., Mirri, S., Salomoni, P., & Delnevo, G. 2026, arXiv preprint arXiv:2601.14791
- [53] Liu, L., Delnevo, G., & Mirri, S. 2023, Journal of Big Data, 10, 31
- [54] Liu, L., Malmert, A., Pouyet, E., Mirri, S., & Delnevo, G. 2025, IEEE Access
- [55] Liu, Q. 2023, Soft Computing, 1
- [56] Liu, Y., Liu, B., Yu, J., Xia, J., & Luo, C. 2022, Applied Artificial Intelligence, 36, 1994232
- [57] Loshchilov, I., & Hutter, F. 2016, arXiv preprint arXiv:1608.03983
- [58] —. 2017, arXiv preprint arXiv:1711.05101
- [59] Lundberg, S. M., & Lee, S.-I. 2017, Advances in neural information processing systems, 30
- [60] Lyons, M. 2021, Journal of Computer Applications in Archaeology, 4
- [61] Ma, J., Peng, Y., Cheng, W., Qiu, M., & Nie, Y. 2021, 213
- [62] Ma, W. 2008, Ma Weidu on Collecting: Ceramic Volum, CCTV 10 Bai jia jiang tan (China Book Council). <https://books.google.it/books?id=QOcU0AEACAAJ>
- [63] Maarroof, K. A. 2017, Research Journal of Applied Sciences, Engineering and Technology, 14, 324
- [64] Mahadevkar, S. V., Khemani, B., Patil, S., et al. 2022, Ieee Access, 10, 107293
- [65] Makridis, M., & Daras, P. 2013, Journal on Computing and Cultural Heritage (JOCCH), 5, 1
- [66] Mao, X. 2010, Studies in the Authentication of Ancient Ceramics (Chinese Academy of Social Sciences Press). <https://books.google.it/books?id=4yvFoQEACAAJ>
- [67] Mazzetto, S. 2024, Heritage, 7, 6432
- [68] McAnany, P. A., & Varela, S. L. L. 1999, Ancient Mesoamerica, 10, 147

- [69] Mendonça, I., Miyaura, M., Fatyanosa, T. N., et al. 2023, *Journal of Archaeological Science : Reports*, 49, 104003
- [70] Merizzi, F., Saillard, P., Acquier, O., et al. 2024, *Heritage Science*, 12, 1
- [71] Ming, Y., Me, R. C., Chen, J. K., & Rahmat, R. W. O. 2023, *Applied Sciences*, 13, 8991
- [72] Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group*, P. 2009, *Annals of internal medicine*, 151, 264
- [73] Mu, T., Wang, F., Wang, X., & Luo, H. 2019, *Ceramics International*, 45, 18140
- [74] Nag, A., & Mishra, S. 2024, in *Review of technologies and disruptive business strategies*, Vol. 3 (Emerald Publishing Limited), 137–162
- [75] Nightingale, A. 2009, *Surgery (Oxford)*, 27, 381
- [76] Niu, C., & Zhang, M. 2022, *Advances in Mathematical Physics*, 2022, 1
- [77] Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. 2021, *International journal of surgery*, 88, 105906
- [78] Paszke, A., Gross, S., Massa, F., et al. 2019, *Advances in neural information processing systems*, 32
- [79] Pawlowicz, L. M., & Downum, C. E. 2021, *Journal of Archaeological Science*, 130, 105375
- [80] Piccoli, C., Aparajeya, P., Papadopoulos, G. T., et al. 2015, *Across Space and Time*, 463
- [81] Portakal, Z. G., Sibel, Oglakci, M., Barad, A., & Topaksu, M. 2024, *Journal of the American Ceramic Society*, 107, 640
- [82] Powers, D. M. 2020, arXiv preprint arXiv:2010.16061
- [83] Prados, & Pavlidis, G. 2025, *Journal of Cultural Heritage Management and Sustainable Development*, 15, 664
- [84] Pratt, W. K. 2007, *Digital image processing: PIKS Scientific inside*, Vol. 4 (Wiley Online Library)
- [85] Rasheed, N. A., & Nordin, M. J. 2015, *Res. J. Appl. Sci. Eng. Technol*, 10, 1396

- [86] Rice, P. M. 2015
- [87] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. 2022, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 10684–10695
- [88] Ruck, L., & Brown, C. T. 2015, Journal of Archaeological Science: Reports, 3, 549
- [89] Ruder, S. 2017, arXiv preprint arXiv:1706.05098
- [90] Ruiz, N., Li, Y., Jampani, V., et al. 2023, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 22500–22510
- [91] Russakovsky, O., Deng, J., Su, H., et al. 2015, International journal of computer vision, 115, 211
- [92] Salimans, T., Goodfellow, I., Zaremba, W., et al. 2016, Advances in neural information processing systems, 29
- [93] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. 2018, in Proceedings of the IEEE conference on computer vision and pattern recognition, 4510–4520
- [94] Schuhmann, C., Beaumont, R., Vencu, R., et al. 2022, Advances in neural information processing systems, 35, 25278
- [95] Selvaraju, R. R., Cogswell, M., Das, A., et al. 2017, in Proceedings of the IEEE international conference on computer vision, 618–626
- [96] Shi, S. 2005, Encyclopedia of Chinese Artefacts Collection and Appreciation: Ceramic, Vol. 2 (Beijing: Beijing Publisher)
- [97] Simonyan, K., & Zisserman, A. 2014, arXiv preprint arXiv:1409.1556
- [98] Sokolova, M., & Lapalme, G. 2009, Information processing & management, 45, 427
- [99] Stanco, F., & Gueli, A. M. 2013, in Digital Photography IX, Vol. 8660, SPIE, 194–203
- [100] Sun, J., Lu, H., Qiao, L., et al. 2023, Ceramics International, 49, 14246
- [101] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. 2016, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2818–2826
- [102] Szegedy, C., Liu, W., Jia, Y., et al. 2015, in Proceedings of the IEEE conference on computer vision and pattern recognition, 1–9

- [103] Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., et al. 2016, IEEE transactions on medical imaging, 35, 1299
- [104] Tan, M., & Le, Q. 2019, in International conference on machine learning, PMLR, 6105–6114
- [105] Teddy, D., Romain, J., Aladine, C., et al. 2015, in 2015 International Conference on Image Processing Theory, Tools and Applications (IPTA), IEEE, 224–228
- [106] Vaswani, A., Shazeer, N., Parmar, N., et al. 2017, Advances in neural information processing systems, 30
- [107] Wang, H. 2019, Identification and Appreciation to Cultural Relics, 15, 70
- [108] Wang, H., Zhang, F., Zhao, M., et al. 2019, in The world wide web conference, 2000–2010
- [109] Wang, K.-g., Geng, G.-h., & Zhou, M.-q. 2008, in 2008 3rd International Conference on Intelligent System and Knowledge Engineering, Vol. 1, IEEE, 699–702
- [110] Weng, Z., Guan, Y., & Luo, H. 2017, JOURNAL OF THE CHINESE CERAMIC SOCIETY, 45, 1833
- [111] Wu, J., Luo, H., Li, J., et al. 2007, Journal of the Chinese Ceramic Society, S1, 39
- [112] Wu, T., Wu, Q., et al. 2023, Mediterranean Archaeology and Archaeometry, 23, 16
- [113] Xie, R., Xia, H., Li, X., & Zhong, C. 2025, in 2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE), IEEE, 1307–1314
- [114] Xu, Y., & Tao, Y. 2022, International Journal of Intangible Heritage, 17, 214
- [115] Yan, X., & Lu, B. 2025, in Proceedings of the 2025 5th International Conference on Internet of Things and Machine Learning, 161–165
- [116] Yang, Y., Wu, H., Yu, D., & Yang, C. 2022, in 2022 International Conference on Culture-Oriented Science and Technology (CoST), IEEE, 6–10
- [117] Yi, J. H., Kang, W., Kim, S.-E., Park, D., & Hong, J.-H. 2021, IEEE Access, 9, 42868
- [118] Yue, X., Chen, J., & Zhuang, Z. 2023, in 2023 IEEE 6th International Conference on Information Systems and Computer Aided Education (ICISCAE), IEEE, 314–318

- [119] Zhang, J., Zhou, Q., Wang, Y., & Luo, H. 2020, Journal of the Chinese Ceramic Society, 48
- [120] Zhang, J., Xiang, X., Xu, B., et al. 2023, Current Opinion in Solid State and Materials Science, 27, 101057
- [121] Zhang, Y., Han, A., Wang, S., et al. 2025, npj Heritage Science, 13, 461
- [122] Zhang, Z., Gong, Z., & Hong, Q. 2021, in The Proceedings of The 8th International Conference on Intelligent Systems and Image Processing
- [123] Zhao, H. 2015, History of Chinese Ceramic Identification (China National Photography Art Publisher). <https://books.google.it/books?id=nivMsgEACAAJ>
- [124] Zhao, H., & He, J. 2025, International Journal of High Speed Electronics and Systems, 2540893
- [125] Zheng, X., Hu, H., & Mingquan, Z. 2012, Journal of Image and Graphics, 17, 1115
- [126] Zhou, J., Lu, Y., Smith, K., et al. 2019, in Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning), 1–8
- [127] Zhou, P., & Wang, K. 2017, in 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), IEEE, 791–797