

**The Egyptian E-Learning University**  
**Faculty of Computers and Information Technology**  
**Master of Software Engineering Program**

# **Ancient Egyptian Statues Classification Methodologies Using Deep Learning**

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master  
of Software Engineering

By  
**Nader Emad Yacoup Hanna**

*Assistant Lecturer, Faculty of Computers and Information Technology, The Egyptian E-Learning  
University*

**Supervised by**

***Dr. Yasser Abdelhamid***

*Associate Professor, Faculty of Graduate  
Studies for Statistical Research, Computer  
Science Department, Cairo University*

***Dr. Rodaina Abdelsalam***

*Assistant Professor, Faculty of Computers and  
Information Technology, The Egyptian E-  
Learning University*

**Egypt – Cairo  
2026**

# Abstract

This thesis investigates artificial intelligence for ancient Egyptian statue recognition through three connected components: a review of deep-learning methods in cultural heritage, the construction of a dedicated benchmark dataset (EgyStatues-v1), and a controlled experimental study of statue classification. The thesis addresses fine-grained, multi-class classification of ancient Egyptian statues from two-dimensional RGB photographs for digital documentation, museum cataloging, and Egyptology-oriented retrieval. To support documented benchmarking, EgyStatues-v1 was curated from publicly accessible museum pages, heritage repositories, and class-guided web-image retrieval, then cleaned through duplicate control and manual verification. The retained benchmark contains 41,161 images across 25 classes and is partitioned into training (28,802), validation (6,163), and test (6,196) sets. Across the reported CNN benchmark, all images were standardized to  $224 \times 224$  pixels and trained under a documented transfer-learning protocol using ImageNet-pretrained frozen backbones and validation-guided early stopping. On the held-out test set, ResNet50V2 achieved the highest accuracy (92.38%), followed closely by EfficientNetV2B0 (92.08%), MobileNetV3Small (90.93%), and VGG19 (83.97%). In addition to overall accuracy, the thesis reports macro-averaged metrics, class-wise analysis, confusion-matrix interpretation, and representative misclassified samples to identify the remaining recognition errors. Exploratory YOLOv11 and YOLOv12 detection runs are discussed as side experiments related to task formulation, while the retained formal benchmark centers on CNN-based transfer-learning classifiers. A supplementary qualitative probe using real-world statue photographs and additional out-of-split web images indicates practical potential, but the claim remains deliberately conservative because the benchmark uses frozen backbones, partly model-specific preprocessing, and qualitative rather than fully quantitative external validation. The principal contribution is a documented benchmark and evaluation framework that can support future work on multimodal heritage AI, partial fine-tuning, transformer-based extensions, and trustworthy museum-facing deployment.

# Acknowledgments

First and foremost, I thank Allah (God) for His countless blessings, guidance, and care throughout this journey. I am deeply grateful to my supervisors, Dr. Yasser Abdelhamid and Dr. Rodaina Abdelsalam, for their continuous support, insightful feedback, and patience throughout this work. I also acknowledge the Egyptian E-Learning University for supporting my master's studies and for providing an academic environment that made this research possible. My sincere thanks go to my colleagues at work, especially Eng. George Sami (Assistant Lecturer), for standing by me and helping me in many situations. I would also like to express my sincere appreciation to Eng. George Sherif for his valuable help, encouragement, and continuous support throughout this journey. Finally, I thank my family for their love, encouragement, and support at every step of this path.

**Signature:** \_\_\_\_\_

# Certification

I certify that this thesis has not been accepted, in substance, for any academic degree and is not being concurrently submitted for any other degree. All sources to which this thesis is indebted are acknowledged, and explicit references are provided.

**Student: Nader Emad Yacoup Hanna**

**Signature:** \_\_\_\_\_

**Date:** \_\_\_\_\_

# List of Publications

1. Emad, N., Abdelhamid, Y., & Abdelsalam, R. (2025, December 8). AI-driven recognition of cultural artifacts: A review of deep learning applications in heritage studies. The 58th Annual International Conference on Data Science, Cairo University, Cairo, Egypt.
2. Emad, N., Abdelhamid, Y., & Abdelsalam, R. (2026). EgyStatues-v1: A curated benchmark dataset for deep learning classification of ancient Egyptian statues. npj Heritage Science. (Under Peer Review)
3. Emad, N., Abdelhamid, Y., & Abdelsalam, R. (2026). Deep learning for ancient Egyptian statue classification. npj Heritage Science. (Under Peer Review)

# Table of Contents

|                                                                                            |    |
|--------------------------------------------------------------------------------------------|----|
| <b><u>Chapter 1 Introduction</u></b>                                                       | 1  |
| <u>1.1 Overview</u>                                                                        | 1  |
| <u>1.2 Problem Statement</u>                                                               | 2  |
| <u>1.3 Thesis Objective</u>                                                                | 3  |
| <u>1.3.1 Research Questions</u>                                                            | 3  |
| <u>1.3.2 Main Contributions</u>                                                            | 4  |
| <u>1.3.3 Scope of the Study</u>                                                            | 4  |
| <u>1.4 Thesis Organization</u>                                                             | 5  |
| <b><u>Chapter 2 Cultural Artifact Recognition: Background and Classical Approaches</u></b> | 6  |
| <u>2.1 Machine Learning and Deep Learning Background</u>                                   | 6  |
| <u>2.1.1 Machine Learning Fundamentals</u>                                                 | 6  |
| <u>2.1.2 Deep Learning Fundamentals</u>                                                    | 8  |
| <u>2.1.3 Supervised Image Classification and Relation to the Thesis</u>                    | 9  |
| <u>2.2 Cultural Heritage Image Recognition Tasks</u>                                       | 10 |
| <u>2.3 Classical (Pre-Deep Learning) Pipelines</u>                                         | 11 |
| <u>2.4 Landmark and Instance Recognition as a Related Problem</u>                          | 11 |
| <u>2.5 Summary of Recent Cultural Heritage Recognition Directions</u>                      | 12 |
| <u>2.5.1 Key Challenges in Heritage Image Recognition</u>                                  | 12 |
| <u>2.5.2 Lessons for Ancient Egyptian Statue Recognition</u>                               | 12 |
| <u>2.6 Dataset Shift, Class Imbalance, and Label Ambiguity</u>                             | 13 |
| <u>2.7 Fine-Grained Visual Categorization of Ancient Egyptian Statues</u>                  | 14 |
| <u>2.8 Implications for Model and Metric Design</u>                                        | 15 |
| <b><u>Chapter 3 Deep Learning Approaches in Cultural Heritage Recognition</u></b>          | 16 |
| <u>3.1 Introduction</u>                                                                    | 16 |
| <u>3.2 CNN and Transfer Learning</u>                                                       | 17 |
| <u>3.3 Generative Models and Data Augmentation</u>                                         | 18 |
| <u>3.4 3D and Multimodal Learning</u>                                                      | 19 |
| <u>3.5 Explainable AI for Trustworthy Heritage Systems</u>                                 | 20 |
| <u>3.6 Detailed Review of Recent Studies</u>                                               | 21 |
| <u>3.6.1 Heritage building and architectural element classification</u>                    | 21 |
| <u>3.6.2 Detection-based heritage documentation</u>                                        | 22 |
| <u>3.6.3 Museum artifact classification and multimodal fusion</u>                          | 22 |
| <u>3.6.4 Landmark recognition and instance-level retrieval</u>                             | 22 |
| <u>3.6.5 Transformers and large vision models in heritage</u>                              | 23 |
| <u>3.6.6 Surveys and bibliometric mappings</u>                                             | 23 |
| <u>3.7 Additional Relevant Studies</u>                                                     | 23 |
| <u>3.8 Comparative Synthesis of the Literature and Methodological Lessons</u>              | 29 |
| <u>3.8.1 Recurring Methodological Limitations</u>                                          | 29 |
| <u>3.8.2 Recurring Strengths in Recent Cultural-Heritage Studies</u>                       | 29 |
| <u>3.8.3 Implications for the Present Thesis</u>                                           | 30 |
| <u>3.9 Research Gaps and Motivation for the Present Thesis</u>                             | 34 |
| <u>3.10 Summary</u>                                                                        | 35 |
| <b><u>Chapter 4 Dataset Construction and Performance Measures</u></b>                      | 36 |

|                                                                                                        |    |
|--------------------------------------------------------------------------------------------------------|----|
| <a href="#"><u>4.1 Introduction</u></a>                                                                | 36 |
| <a href="#"><u>4.2 Dataset Construction</u></a>                                                        | 36 |
| <a href="#"><u>4.2.1 Related Cultural Heritage Datasets (Brief Overview)</u></a>                       | 36 |
| <a href="#"><u>4.2.2 The Proposed EgyStatues-v1 Dataset</u></a>                                        | 40 |
| <a href="#"><u>4.2.2.1 Data Acquisition via Automated Web Scraping</u></a>                             | 41 |
| <a href="#"><u>4.2.2.2 Manual Cleaning and Curation</u></a>                                            | 42 |
| <a href="#"><u>4.2.2.3 Data Augmentation for Minority Classes</u></a>                                  | 44 |
| <a href="#"><u>4.2.2.4 Dataset Packaging and Versioning</u></a>                                        | 45 |
| <a href="#"><u>4.2.2.5 Dataset Split and Class Distribution</u></a>                                    | 46 |
| <a href="#"><u>4.2.2.6 Data and Code Availability</u></a>                                              | 50 |
| <a href="#"><u>4.2.2.7 Data Ethics, Provenance, and Responsible Reuse</u></a>                          | 50 |
| <a href="#"><u>4.3 Benchmark Dataset Evaluation</u></a>                                                | 51 |
| <a href="#"><u>4.3.1 Strengths of the Proposed Benchmark</u></a>                                       | 51 |
| <a href="#"><u>4.3.2 Remaining Limitations</u></a>                                                     | 51 |
| <a href="#"><u>4.4 Performance Measures</u></a>                                                        | 52 |
| <a href="#"><u>4.4.1 Common Performance Measures</u></a>                                               | 52 |
| <a href="#"><u>4.4.2 Other Measures and Diagnostic Tools</u></a>                                       | 54 |
| <a href="#"><u>4.4.3 Detection Measures Used for YOLO Baselines</u></a>                                | 54 |
| <a href="#"><u>4.5 Summary</u></a>                                                                     | 54 |
| <b><a href="#"><u>Chapter 5 Proposed Approach for Ancient Egyptian Statue Classification</u></a></b>   | 55 |
| <a href="#"><u>5.1 Introduction</u></a>                                                                | 55 |
| <a href="#"><u>5.2 Input Dataset Preprocessing</u></a>                                                 | 56 |
| <a href="#"><u>5.2.1 Rationale for Input Standardization</u></a>                                       | 56 |
| <a href="#"><u>5.2.2 Label Consistency and Split Integrity</u></a>                                     | 56 |
| <a href="#"><u>5.2.3 Augmentation Strategy</u></a>                                                     | 57 |
| <a href="#"><u>5.3 Selection of Candidate Models</u></a>                                               | 57 |
| <a href="#"><u>5.3.1 Architecture-Selection Rationale</u></a>                                          | 58 |
| <a href="#"><u>5.4 Training Configuration</u></a>                                                      | 58 |
| <a href="#"><u>5.4.1 Optimization Strategy and Model Selection</u></a>                                 | 59 |
| <a href="#"><u>5.4.2 Reproducibility and Practical Deployment</u></a>                                  | 59 |
| <a href="#"><u>5.4.3 Deployment Threat Model and Practical Constraints</u></a>                         | 60 |
| <a href="#"><u>5.4.4 Model-Selection Criteria for Heritage Deployment</u></a>                          | 61 |
| <a href="#"><u>5.4.5 Fair-Comparison Protocol and Reproducibility Safeguards</u></a>                   | 62 |
| <a href="#"><u>5.5 Summary</u></a>                                                                     | 63 |
| <b><a href="#"><u>Chapter 6 Experimental Work</u></a></b>                                              | 64 |
| <a href="#"><u>6.1 Introduction</u></a>                                                                | 64 |
| <a href="#"><u>6.2 Hardware and Software Settings</u></a>                                              | 64 |
| <a href="#"><u>6.3 Dataset Used in the Experiments</u></a>                                             | 66 |
| <a href="#"><u>6.4 Model Benchmarking</u></a>                                                          | 66 |
| <a href="#"><u>6.5 Results and Discussion</u></a>                                                      | 68 |
| <a href="#"><u>6.5.1 EfficientNetV2B0 Results</u></a>                                                  | 68 |
| <a href="#"><u>6.5.2 ResNet50V2 Results</u></a>                                                        | 70 |
| <a href="#"><u>6.5.3 MobileNetV3Small Results</u></a>                                                  | 72 |
| <a href="#"><u>6.5.4 VGG19 Results</u></a>                                                             | 74 |
| <a href="#"><u>6.5.5 Why Image Classification Was More Suitable Than Detection in This Setting</u></a> | 76 |
| <a href="#"><u>6.5.6 YOLOv12 Results (Detection Baseline)</u></a>                                      | 77 |

|                                                                                           |    |
|-------------------------------------------------------------------------------------------|----|
| <a href="#"><u>6.5.7 YOLOv11 Results (Detection Baseline)</u></a>                         | 79 |
| <a href="#"><u>6.5.8 Per-class Analysis and Relation to Data Distribution</u></a>         | 81 |
| <a href="#"><u>6.5.9 Relation to the Broader Literature</u></a>                           | 84 |
| <a href="#"><u>6.5.10 Practical Implications for Cataloging and Museum Use</u></a>        | 85 |
| <a href="#"><u>6.5.11 Threats to Validity and Remaining Limitations</u></a>               | 85 |
| <a href="#"><u>6.5.12 Recommended Human-in-the-Loop Workflow</u></a>                      | 86 |
| <a href="#"><u>6.5.13 Robustness, External Validity, and Generalization</u></a>           | 87 |
| <a href="#"><u>6.5.14 Error Analysis and Plausible Sources of Confusion</u></a>           | 87 |
| <a href="#"><u>6.6 Summary</u></a>                                                        | 88 |
| <b><a href="#"><u>Chapter 7 Conclusion and Future Work</u></a></b>                        | 91 |
| <a href="#"><u>7.1 Thesis Contributions</u></a>                                           | 91 |
| <a href="#"><u>7.2 Future Work</u></a>                                                    | 92 |
| <a href="#"><u>7.3 Strategic Roadmap for Future Benchmark Releases and Deployment</u></a> | 93 |
| <b><a href="#"><u>References</u></a></b>                                                  | 94 |

# List of Figures

|                                                                                                                           |    |
|---------------------------------------------------------------------------------------------------------------------------|----|
| Figure 4.1: Overview of the data acquisition and preparation pipeline for EgyStatues-v1.                                  | 43 |
| Figure 4.2: Examples of augmented samples generated using ImageDataGenerator (rotations, shifts, shear, and zoom).        | 45 |
| Figure 4.3: Number of images in each dataset split (training/validation/test).                                            | 48 |
| Figure 4.4: Total number of images per class in EgyStatues-v1 (sorted).                                                   | 49 |
| Figure 5.1: Overview of the proposed end-to-end pipeline for ancient Egyptian statue classification.                      | 55 |
| Figure 6.1: Comparative test accuracies of classification models on EgyStatues-v1.                                        | 67 |
| Figure 6.2: EfficientNetV2B0 training and validation loss across epochs.                                                  | 69 |
| Figure 6.3: EfficientNetV2B0 training and validation accuracy across epochs.                                              | 69 |
| Figure 6.4: Qualitative test-set predictions of EfficientNetV2B0 (random samples with true vs predicted labels).          | 70 |
| Figure 6.5: Single-image inference example using EfficientNetV2B0 (web image predicted as Statue of Anubis).              | 70 |
| Figure 6.6: ResNet50V2 training and validation loss across epochs.                                                        | 71 |
| Figure 6.7: ResNet50V2 training and validation accuracy across epochs.                                                    | 71 |
| Figure 6.8: Qualitative test-set predictions of ResNet50V2 (random samples with true vs predicted labels).                | 72 |
| Figure 6.9: Single-image inference examples using ResNet50V2 (web images predicted as Khufu Statue and Statue of Anubis). | 72 |
| Figure 6.10: MobileNetV3Small training and validation loss across epochs.                                                 | 73 |
| Figure 6.11: MobileNetV3Small training and validation accuracy across epochs.                                             | 73 |
| Figure 6.12: Qualitative test-set predictions of MobileNetV3Small (random samples with true vs predicted labels).         | 74 |
| Figure 6.13: Single-image inference example using MobileNetV3Small (web image predicted as Statue of Anubis).             | 74 |
| Figure 6.14: VGG19 training and validation loss/accuracy across epochs.                                                   | 75 |
| Figure 6.15: Qualitative test-set predictions of VGG19 (random samples with true vs predicted labels).                    | 75 |
| Figure 6.16: Single-image inference examples using VGG19 (web images predicted as Seneb and Nefertiti Bust).              | 76 |
| Figure 6.17: YOLOv12 overall detection metrics on the validation split.                                                   | 78 |
| Figure 6.18: Per-class mAP@0.5 for YOLOv12 on the validation split.                                                       | 79 |
| Figure 6.19: Comparison of YOLOv11 vs YOLOv12 overall detection metrics on the validation split.                          | 80 |
| Figure 6.20: Per-class mAP@0.5 for YOLOv11 on the validation split.                                                       | 81 |
| Figure 6.21: Representative misclassified test samples for ResNet50V2.                                                    | 83 |
| Figure 6.22: Confusion matrix of ResNet50V2 on the held-out test split.                                                   | 84 |

# List of Tables

|                                                                                                                                        |    |
|----------------------------------------------------------------------------------------------------------------------------------------|----|
| Table 3.1: Consolidated comparison of representative cultural-heritage deep-learning studies reviewed in this chapter.                 | 31 |
| Table 3.2: Main research gaps identified in the cultural-heritage literature and the corresponding contribution of the present thesis. | 34 |
| Table 4.1: Comparison of representative cultural-heritage datasets related to EgyStatues-v1.                                           | 37 |
| Table 4.2: The 25 statue categories included in EgyStatues-v1.                                                                         | 45 |
| Table 4.3: Class-wise distribution across training, validation, and test splits.                                                       | 47 |
| Table 5.1: Rationale for selecting the compared architectures in the proposed benchmarking pipeline.                                   | 57 |
| Table 5.2: Summary of training configuration and hyperparameters for all evaluated models.                                             | 60 |
| Table 5.3: Deployment-oriented criteria used to interpret the model families beyond raw accuracy.                                      | 62 |
| Table 6.1: Comparative performance of classification models on the test split (accuracy).                                              | 67 |
| Table 6.2: Overall detection performance of YOLO baselines on the validation split.                                                    | 78 |
| Table 6.3: Five lowest-performing classes under YOLOv12 by mAP@0.5 (validation split).                                                 | 78 |
| Table 6.4: The five lowest-performing YOLOv11 classes on the validation split, ranked from the lowest to the highest mAP@0.5.          | 80 |
| Table 6.5: Macro-averaged precision, recall, and F1-score of the classification models on the held-out test split.                     | 82 |
| Table 6.6: Operational interpretation of common error types and recommended mitigation actions.                                        | 88 |
| Table 6.7: Suggested model families for different practical deployment scenarios derived from the experimental findings.               | 89 |

# List of Abbreviations

|                 |                                            |
|-----------------|--------------------------------------------|
| <b>AI</b>       | Artificial Intelligence                    |
| <b>BN</b>       | Batch Normalization                        |
| <b>BoVW</b>     | Bag of Visual Words                        |
| <b>CH</b>       | Cultural Heritage                          |
| <b>CNN</b>      | Convolutional Neural Network               |
| <b>CSV</b>      | Comma-Separated Values                     |
| <b>DL</b>       | Deep Learning                              |
| <b>FN</b>       | False Negative                             |
| <b>FP</b>       | False Positive                             |
| <b>GAN</b>      | Generative Adversarial Network             |
| <b>GAP</b>      | Global Average Pooling                     |
| <b>GLDv2</b>    | Google Landmarks Dataset v2                |
| <b>GPU</b>      | Graphics Processing Unit                   |
| <b>Grad-CAM</b> | Gradient-weighted Class Activation Mapping |
| <b>IoU</b>      | Intersection over Union                    |
| <b>kNN</b>      | k-Nearest Neighbors                        |
| <b>mAP</b>      | Mean Average Precision                     |
| <b>ML</b>       | Machine Learning                           |
| <b>RAM</b>      | Random Access Memory                       |
| <b>RGB</b>      | Red, Green, and Blue Image Channels        |
| <b>SIFT</b>     | Scale-Invariant Feature Transform          |
| <b>SURF</b>     | Speeded-Up Robust Features                 |
| <b>SVM</b>      | Support Vector Machine                     |
| <b>TL</b>       | Transfer Learning                          |
| <b>TN</b>       | True Negative                              |
| <b>TP</b>       | True Positive                              |
| <b>URL</b>      | Uniform Resource Locator                   |
| <b>ViT</b>      | Vision Transformer                         |
| <b>XAI</b>      | Explainable Artificial Intelligence        |
| <b>YOLO</b>     | You Only Look Once                         |

# Chapter 1

## Introduction

### 1.1 Overview

Ancient Egyptian statues are among the most recognizable and historically significant cultural artifacts in the world. They support archaeological documentation, museum curation, digital heritage access, and Egyptological interpretation because they encode information through iconography, posture, material, inscriptions, and stylistic conventions. As museum collections, public repositories, and web-accessible archives continue to expand, the ability to organize and retrieve statue imagery at scale becomes increasingly important. Manual classification, however, remains time-consuming and can vary across specialists when photographs are cropped, degraded, poorly lit, or drawn from visually similar iconographic families. Recent progress in deep learning, therefore, creates an opportunity to support expert workflows through visual models that learn discriminative patterns directly from image data [Fiorucci et al., 2020] [Mishra and Lourenço, 2024].

In a digital-heritage setting, automatic recognition should support rather than replace archaeologists, curators, and Egyptologists. Its practical value lies in accelerating indexing, assisting search within large archives, enriching metadata suggestions, surfacing visually related objects, and supporting educational interfaces in museums and online collections. Even when a classifier performs well, however, it remains one component within a broader documentation workflow that depends on provenance, chronology, material analysis, conservation context, and expert interpretation [Fiorucci et al., 2020] [Paolanti et al., 2026].

Ancient Egyptian statuary also poses a demanding computer-vision problem because the categories are often fine-grained rather than visually distant. Many statue classes share frontal poses, similar silhouettes, related stone colors, museum-display backgrounds, and repeated iconographic motifs. The classifier may need to distinguish a crown, headdress, animal association, seated posture, pedestal relation, or small inscription region that occupies

only a limited part of the photograph. In addition, web and museum photographs vary in scale, viewpoint, cropping, reflections, restoration state, image quality, and background clutter. These conditions justify a controlled benchmark design, since model performance cannot be interpreted reliably without fixed data splits, explicit class definitions, and careful discussion of residual bias [Winterbottom et al., 2022] [Siountri and Anagnostopoulos, 2023].

## 1.2 Problem Statement

This thesis addresses the problem of fine-grained multi-class classification of ancient Egyptian statues from 2D RGB photographs. Given an input image, the goal is to predict one of 25 statue categories. The problem is challenging due to intra-class variation (pose, viewpoint, lighting, background), inter-class similarity (shared materials and stylistic motifs), and the limited availability of publicly documented, statue-specific benchmark datasets.

Formally, the task can be described as learning a mapping from an RGB image to one label among twenty-five statue categories. However, the practical challenge is richer than the formal definition suggests. A model must generalize across museum photos, catalog images, photographs from heritage websites, and informal web images, while still attending to fine-grained cues such as silhouette, headdress, posture, facial structure, inscriptions, and local decorative details.

A further difficulty is the distinction between category-level classification and instance-level heritage recognition. Many cultural-heritage systems aim to identify a specific object, monument, or artifact instance. The present thesis instead targets category recognition; nevertheless, both task types are affected by viewpoint change, illumination shift, background bias, and incomplete object visibility. This connection positions the thesis within the wider family of heritage-vision problems [Weyand et al., 2020] [Winterbottom et al., 2022].

## 1.3 Thesis Objective

The main objective of this thesis is to develop a comprehensive and reproducible framework for the classification of ancient Egyptian statues using deep-learning techniques. This objective is addressed through four complementary goals covering dataset construction, documented data preparation, model benchmarking, and limitation-aware experimental evaluation.

First, the thesis aims to construct a large and carefully curated dataset dedicated to ancient Egyptian statue categories, referred to as EgyStatues-v1. The dataset is designed to provide a reliable benchmark for training and evaluating image-classification models in this specialized cultural-heritage domain.

Second, the thesis seeks to develop and document a preprocessing and augmentation workflow suitable for heterogeneous statue imagery. This pipeline is intended to reduce the effects of variations in image size, viewpoint, illumination, background, and class representation.

Third, the thesis aims to benchmark modern deep-learning architectures and determine which pretrained backbone provides the strongest balance between classification accuracy and practical usability on the proposed dataset.

Finally, the thesis seeks to establish a reproducible benchmarking workflow that integrates dataset versioning, documented training configurations, limitation-aware interpretation, and defensible evaluation practices for cultural-heritage artificial intelligence applications.

### 1.3.1 Research Questions

The work is guided by four main research questions. RQ1: Can a statue-specific benchmark for ancient Egyptian categories be constructed at a scale and quality sufficient for reproducible deep-learning comparison? RQ2: Which methodological lessons from the broader cultural-heritage AI literature are most relevant to this problem, particularly with respect to

data scarcity, transfer learning, interpretability, and deployment realism? RQ3: Among the compared CNN-based transfer-learning backbones, which model offers the strongest balance between recognition quality and practical deployability for this domain? RQ4: What limitations remain before such a system can be treated as a trustworthy assistive component in museum-oriented cataloging and documentation workflows?

### **1.3.2 Main Contributions**

The thesis makes four principal contributions. First, it synthesizes recent deep-learning directions in cultural heritage and identifies the methodological gaps most relevant to ancient Egyptian statue recognition. Second, it introduces and documents a curated benchmark dataset focused on twenty-five statue categories. Third, it reports a controlled CNN-centered comparison, interprets class-wise behavior and misclassifications, and distinguishes what the results support from what they do not yet justify in real heritage workflows. Fourth, from a software-engineering perspective, it establishes a reproducible benchmark workflow built around versioned data packaging, explicit training configurations, diagnostic evaluation, and cautious deployment-oriented interpretation. This framing is important because cultural-heritage AI projects are less useful when they report only leaderboard metrics while neglecting benchmark design, interpretability, and deployment constraints.

### **1.3.3 Scope of the Study**

The current work focuses on image-based category classification from two-dimensional photographs. It does not claim to solve chronology estimation, provenance verification, forgery detection, inscription understanding, or full semantic cataloging. Likewise, although detection baselines and future transformer extensions are discussed, the retained formal benchmark of the main experimental contribution is intentionally limited to CNN-based transfer-learning classifiers so that the scope remains interpretable and consistent with the available archived experiments. Making this scope explicit strengthens the thesis because it clarifies what the proposed method can legitimately achieve and what remains to be addressed by future multimodal, 3D-aware, or expert-assisted systems.

## 1.4 Thesis Organization

This thesis is organized as follows. Chapter 2 establishes the machine-learning, deep-learning, and cultural-heritage recognition background, including classical approaches and the challenges of fine-grained statue categorization. Chapter 3 reviews deep-learning approaches in cultural heritage, including transfer learning, augmentation, multimodal learning, explainability, and recent methodological trends. Chapter 4 presents the construction, curation, split organization, and evaluation measures of the proposed EgyStatues-v1 benchmark. Chapter 5 details the proposed statue-classification framework and the reasoning behind the selected model families and training setup. Chapter 6 presents the experimental configuration, benchmark results, diagnostic analysis, and deployment-oriented interpretation. Chapter 7 concludes the thesis and outlines future work.

The organization reflects three connected components of the research. The literature-review chapters establish the conceptual and methodological background, the dataset-construction chapter presents EgyStatues-v1 and its curation process, and the methodology and experimental chapters describe the proposed framework and controlled empirical evaluation. Together, these components form an integrated research narrative that moves from field analysis to dataset creation and comparative benchmarking.

## **Chapter 2**

# **Cultural Artifact Recognition: Background and Classical Approaches**

### **2.1 Machine Learning and Deep Learning Background**

Artificial intelligence is a broad field concerned with the development of computational systems capable of performing tasks that normally require human intelligence, such as perception, reasoning, prediction, and decision-making. Machine learning is a major subfield of artificial intelligence in which computational models learn patterns from data instead of relying exclusively on manually programmed rules. Deep learning is, in turn, a specialized branch of machine learning that uses neural networks with multiple computational layers to learn increasingly abstract data representations [Goodfellow et al., 2016].

The distinction between these concepts is important for the present thesis. Artificial intelligence represents the overall research domain, machine learning provides the general data-driven learning framework, and deep learning provides the multilayer neural architectures used for visual feature learning. The statue-classification problem addressed in this thesis is therefore an artificial-intelligence application formulated as a supervised machine-learning task and implemented using deep convolutional neural networks.

#### **2.1.1 Machine Learning Fundamentals**

Machine learning is a branch of artificial intelligence concerned with developing computational models that learn useful patterns from data and use those patterns to make predictions or decisions on previously unseen examples. Unlike rule-based systems, in which the required behavior is explicitly programmed, machine-learning systems improve their behavior by analyzing representative data. The quality of the learned model therefore depends on several factors, including the amount and reliability of the available data, the suitability of

the selected features, the learning algorithm, and the evaluation procedure [Goodfellow et al., 2016].

Machine-learning methods are commonly organized into supervised, unsupervised, semi-supervised, and reinforcement-learning paradigms. Supervised learning uses labeled examples in which each input is associated with a known target output. Its common applications include classification, where the target is a discrete category, and regression, where the target is a continuous numerical value. Unsupervised learning analyzes unlabeled observations to identify hidden structures or relationships, such as clusters and lower-dimensional representations. Semi-supervised learning combines a relatively small labeled subset with a larger collection of unlabeled data, whereas reinforcement learning enables an agent to learn through interaction with an environment and feedback based on its actions [Goodfellow et al., 2016].

The problem addressed in this thesis is formulated as supervised multi-class image classification. Each photograph is associated with one of twenty-five predefined ancient Egyptian statue categories, and the learning process uses these labeled examples to distinguish the visual characteristics associated with each class. After training, the classifier receives a previously unseen photograph and produces a predicted class label. This setting differs from binary classification, which distinguishes between only two alternatives, and from regression, which predicts a continuous value rather than a semantic category.

A supervised-learning workflow typically separates the available examples into training, validation, and test partitions. The training data are used to learn the model parameters. The validation data are used to monitor performance during development, guide hyperparameter choices, support early stopping, and select the retained model configuration. The test data are then used to estimate performance on examples that were not used for parameter learning or model selection. Maintaining this separation is essential because evaluating repeatedly on the test set can unintentionally influence model-development decisions and produce an overly optimistic estimate of generalization.

Generalization refers to the ability of a learned model to perform successfully on unseen

examples rather than merely reproducing patterns from its training data. A model may show low training error while performing less effectively on validation or test images, a condition commonly described as overfitting. Overfitting is particularly relevant to cultural-heritage imagery because datasets may contain repeated viewpoints, source-specific backgrounds, uneven class representation, and limited visual diversity. Appropriate data partitioning, validation-guided model selection, regularization, augmentation, and careful error analysis are therefore important components of a reliable machine-learning workflow.

Traditional machine-learning approaches to image recognition normally separate feature extraction from classification. Visual properties are first represented using manually designed descriptors, such as color, texture, edges, shapes, SIFT, SURF, or bag-of-visual-words features. These representations are then provided to classifiers such as support-vector machines, k-nearest neighbors, decision trees, or random forests. Such methods can remain effective when datasets are limited, and the relevant visual characteristics are well understood. However, their dependence on manually selected representations can restrict their ability to capture the complex and fine-grained visual variations present in ancient Egyptian statue photographs.

This limitation motivates the transition from traditional machine learning to deep learning. Instead of relying primarily on predefined image descriptors, deep-learning models can learn hierarchical visual representations directly from the training images. The following subsection introduces the fundamental principles of deep learning and explains why convolutional neural networks are particularly appropriate for the image-classification problem considered in this thesis.

### **2.1.2 Deep Learning Fundamentals**

Deep learning reduces the dependence on manually engineered features by learning hierarchical representations directly from data. A deep neural network is composed of multiple layers of interconnected computational units. Each layer transforms the representation produced by the preceding layer using learned weights, biases, and nonlinear activation functions. During training, the prediction error is propagated backward through the network,

and an optimization algorithm updates the model parameters to reduce the loss [Goodfellow et al., 2016].

For image-analysis tasks, Convolutional Neural Networks are among the most widely used deep-learning architectures. Convolutional layers apply learnable filters across local image regions and preserve spatial relationships among visual elements. Early layers commonly respond to relatively simple patterns such as edges, orientations, and textures, while deeper layers can represent more complex shapes, object parts, and class-discriminative structures. Pooling or down sampling operations reduce spatial dimensions and improve computational efficiency, while the final classification layers transform the learned representation into class scores or probabilities.

This hierarchical feature-learning ability is particularly relevant to ancient Egyptian statue classification. Many categories share similar materials, frontal poses, museum backgrounds, and global silhouettes. Consequently, successful classification may depend on relatively small visual cues, including headdress geometry, crown shape, associated animals, facial proportions, seated or standing posture, and the relationship between the statue and its pedestal. Deep neural networks can learn combinations of these local and global patterns without requiring every discriminative characteristic to be specified manually.

Deep-learning models nevertheless remain sensitive to the quantity and quality of their training data. When a network learns patterns that are highly specific to the training examples but do not generalize to unseen images, overfitting occurs. Data augmentation, regularization, early stopping, careful validation, and transfer learning are commonly used to reduce this risk. The detailed application of these techniques in cultural-heritage research is discussed in Chapter 3, while their implementation in the proposed classification framework is described in Chapters 5 and 6.

### **2.1.3 Supervised Image Classification and Relation to the Thesis**

Image classification assigns a semantic label to an entire input image. This differs from object detection, which additionally estimates the location of one or more objects, and from

instance recognition, which attempts to identify the exact physical object represented in an image. The primary task of this thesis is category-level image classification: each input photograph is assigned to one of twenty-five ancient Egyptian statue categories.

The retained experimental framework applies transfer learning using ImageNet-pretrained CNN backbones. Rather than training all convolutional parameters from random initialization, the pretrained networks provide previously learned visual representations, while a task-specific classification head maps those representations to the statue categories. Under the retained benchmark setting, the pretrained backbones are used as frozen feature extractors, allowing the comparison to focus on how effectively different architecture families represent the proposed dataset under controlled conditions.

This section establishes the general conceptual foundation required for the remainder of the thesis. The following sections introduce cultural-heritage image-recognition tasks, review classical pre-deep-learning pipelines, and discuss the dataset and recognition challenges that motivate the use of modern deep-learning models. Chapter 3 then provides a more detailed review of deep-learning applications, transfer learning, augmentation, multimodal methods, and explainability in cultural-heritage research.

## **2.2 Cultural Heritage Image Recognition Tasks**

Cultural heritage (CH) visual analytics spans multiple computer-vision tasks, including (i) object/category classification of artifacts, (ii) instance-level recognition and landmark recognition, (iii) object detection and segmentation for condition assessment, and (iv) retrieval and re-identification for museum inventory and visitor applications. In practice, the same artifact may appear under different lighting, viewpoints, occlusions, and background contexts, which motivates robust feature representations and careful dataset curation [Fiorucci et al., 2020] [Mishra and Lourenço, 2024].

In practical heritage systems, these tasks rarely exist in isolation. A museum platform may begin with image classification to assign a coarse label, continue with retrieval to surface similar objects, and finally rely on metadata or expert review to confirm identity.

Understanding this workflow is important because it shows why a statue classifier can still be valuable even if it does not solve full instance-level identification [Rei et al., 2023] [Wen et al., 2024].

### 2.3 Classical (Pre-Deep Learning) Pipelines

Before deep learning, CH recognition typically relied on hand-crafted local descriptors (e.g., SIFT/SURF), bag-of-visual-words (BoVW), spatial pyramids, and classical classifiers (SVM, kNN, Random Forest). These pipelines can be effective with limited data, but their performance degrades on fine-grained categories where discriminative cues are subtle (material texture, iconographic details) and when images exhibit strong viewpoint changes [Fiorucci et al., 2020].

Classical pipelines nevertheless established several principles that remain useful today. They emphasized local detail, robustness to geometric variation, and careful feature engineering under limited data. In heritage analysis, this mattered because many objects share a common global form while differing in localized patterns such as ornament, edge profiles, or inscription regions. Modern CNNs learn such representations automatically, but the underlying need for discriminative local evidence has not disappeared [Fiorucci et al., 2020].

Another lasting lesson from the classical era is the value of retrieval-oriented thinking. Instead of asking only whether an image belongs to a class, retrieval methods ask which archived images it most resembles. This perspective is especially relevant for statues and museum artifacts, where visual neighbors may reveal class ambiguity, mislabeled samples, or meaningful sub-groups within a broad category [Weyand et al., 2020].

### 2.4 Landmark and Instance Recognition as a Related Problem

Landmark recognition benchmarks provide a close analogue to statue identification because they involve fine-grained recognition under real-world capture conditions. The Google Landmarks Dataset v2 (GLDv2) is a large-scale benchmark for instance-level recognition and retrieval, and it highlights challenges such as scale, occlusion, and domain shift. In this thesis, we do not train on GLDv2; however, its problem formulation and evaluation culture inform

the design of robust evaluation protocols for heritage applications.

Winterbottom et al. [2022] further strengthen this connection by showing that deep networks can perform exact-object recognition on a large museum-artifact dataset under realistic image variation. Their work is not about Egyptian statue categories specifically, but it demonstrates that cultural-heritage imagery is amenable to high-capacity visual learning when enough curated examples are available. It also reinforces the importance of transfer learning and careful dataset design in heritage domains.

## **2.5 Summary of Recent Cultural Heritage Recognition Directions**

Recent CH research increasingly adopts transfer learning, data augmentation, and lightweight architectures for deployable solutions. Work also expands toward transformers, multimodal fusion (image + metadata), and generative augmentation. These trends motivate the selection of modern CNN backbones (EfficientNetV2B0 [Tan and Le, 2021], ResNet50V2 [He et al., 2016], MobileNetV3Small [Howard et al., 2019], VGG19 [Simonyan and Zisserman, 2015]) and the inclusion of YOLO detectors as baselines for scenarios where statues appear in complex scenes [Fiorucci et al., 2020] [Xu et al., 2025] [Lu et al., 2026].

### **2.5.1 Key Challenges in Heritage Image Recognition**

Across the literature, four challenges recur. The first is data scarcity: specialist classes may have few reliable images, which makes deep models sensitive to overfitting. The second is domain shift between controlled documentation photographs and uncontrolled web images. The third is semantic ambiguity, where objects share style, material, or iconographic conventions. The fourth is the need for transparency, since heritage professionals often require interpretable evidence rather than purely opaque predictions [Fiorucci et al., 2020] [Paolanti et al., 2026].

### **2.5.2 Lessons for Ancient Egyptian Statue Recognition**

For the present thesis, these challenges mean that model choice alone is insufficient. The study also requires coherent labels, a training strategy that can tolerate moderate imbalance

and real-world noise, and an evaluation protocol that separates raw performance from practical usability. The thesis therefore combines dataset construction, architecture comparison, and critical discussion instead of treating the task as a narrow leaderboard problem [Wisniamongkol et al., 2024] [Paolanti et al., 2026].

Ancient Egyptian statue recognition is therefore treated here as a fine-grained cultural-heritage classification problem under heterogeneous image conditions. It sits at the intersection of benchmark design, visual representation learning, curation quality, and domain-aware interpretation. This framing explains the emphasis on dataset construction and experimental fairness, rather than on architecture novelty alone. It also clarifies the comparison between classification and detection: the most suitable model family depends not only on raw performance, but also on the task structure and intended use scenario [Winterbottom et al., 2022] [Mishra and Lourenço, 2024].

These challenges lead to three methodological implications. First, the benchmark should prioritize architectures that can preserve both local and global information, which justifies the use of transfer-learned CNN backbones and motivates comparison with additional paradigms rather than assuming a single family will suffice. Second, evaluation should not stop at one headline metric. Accuracy is important for the classification task, but learning-curve behavior, qualitative predictions, and class-wise variation provide a more faithful picture of what the model has actually learned. Third, interpretability and responsible deployment should be considered early. In heritage contexts, a high-performing model is still incomplete if it offers no support for curator trust, error inspection, or human correction [Paolanti et al., 2026].

## **2.6 Dataset Shift, Class Imbalance, and Label Ambiguity**

Dataset shift presents an important challenge. Images collected from the open web rarely follow a single acquisition protocol. They may include museum photographs, catalog scans, tourist photos, edited backgrounds, low-resolution thumbnails, collages, and images with substantial compression artifacts. This introduces appearance shifts that are not purely semantic. The same class can be seen under dramatically different illumination conditions, scales, and backgrounds, while two different classes can appear deceptively similar when

photographed in controlled museum displays. Dataset shift is therefore not a minor nuisance; it directly affects the meaning of the reported accuracy because success on one source distribution does not guarantee success on another [Mishra and Lourenço, 2024] [Boutet and Hallot, 2025].

Class imbalance represents a second important challenge. Even when the final benchmark is numerically more balanced than the raw web scrape, effective difficulty remains unequal across classes. Some classes are visually prototypical and easy to recognize. Others include sub-variants, damaged samples, or ambiguous viewpoints that reduce separability. This is why aggregate accuracy should be interpreted alongside per-class behavior and qualitative inspection. A benchmark can appear balanced in counts while still being imbalanced in representation quality, pose diversity, or iconographic complexity. Recent heritage literature increasingly emphasizes this distinction, especially when datasets are curated from heterogeneous public sources rather than from a controlled acquisition campaign [Mishra and Lourenço, 2024] [Colace et al., 2025].

Label ambiguity further complicates the problem. Heritage categories often inherit naming conventions from museum practice, archaeological convention, popular online usage, or educational resources, and these conventions do not always align. A single web image may be tagged using a broad label, a specific artifact name, a dynasty-oriented description, or a phrase unrelated to the benchmark taxonomy. For a statue dataset, this matters because visually similar royal, divine, or sphinx-related objects can be mislabeled online even when the photograph itself is clear. Curation and class-definition discipline are therefore as important as model selection. Ambiguity should be reduced not only by deleting noisy samples but also by documenting inclusion rules, exclusion criteria, and edge cases. This thesis adopts that perspective by treating the dataset as a scholarly contribution rather than a technical by-product [Colace et al., 2025] [Paolanti et al., 2026].

## **2.7 Fine-Grained Visual Categorization of Ancient Egyptian Statues**

The literature in cultural heritage repeatedly shows that visually similar classes, limited data, and inconsistent acquisition settings remain central obstacles. Winterbottom et al. [2022]

demonstrated that even when the task is shifted toward antiquities and illicit-trafficking scenarios, careful feature learning is still needed because heritage objects are strongly affected by viewpoint, wear, occlusion, and collection heterogeneity. Similar concerns appear in studies of architectural-heritage recognition, where models must distinguish classes that are semantically meaningful but visually overlapping in color, texture, and geometry [Siountri and Anagnostopoulos, 2023] [Ottoni and Ottoni, 2025]. These findings are directly relevant here: the difficulty of the proposed benchmark does not arise from object absence, but from the closeness of the classes and the variability of real images.

Fine-grained recognition is more demanding than generic image classification because the target categories share much of their visual vocabulary. In this thesis, many classes belong to the same broad visual family: carved statues, often photographed frontally, frequently preserved in museum environments, and sometimes captured under similar lighting and background conditions. The classifier must therefore rely on subtle cues such as crown shape, posture, seated versus standing composition, associated animals, inscriptions, facial proportions, headdress geometry, and the relation between the figure and its pedestal. These cues are meaningful from an art-historical perspective, but they may occupy only a small region of the image. Transfer learning is useful in this setting because pretrained backbones provide robust low- and mid-level visual filters, while the final classifier learns details that separate one heritage category from another [Belhi et al., 2021] [Fiorucci et al., 2020].

## **2.8 Implications for Model and Metric Design**

Taken together, the discussion in this chapter implies that benchmark design, class definition, and deployment context must be treated as first-class methodological variables rather than secondary implementation details. Accordingly, the present thesis evaluates models not only by headline accuracy, but also by their suitability for fine-grained recognition, their sensitivity to heterogeneous image conditions, and their practical value for museum-oriented classification workflows.

# **Chapter 3**

## **Deep Learning Approaches in Cultural Heritage Recognition**

### **3.1 Introduction**

Deep learning has changed cultural-heritage analysis by reducing dependence on manually engineered descriptors and enabling pretrained visual backbones to be adapted to a wide range of heritage-related tasks. The cultural-heritage domain, however, is not simply a smaller version of generic computer vision. It presents distinctive constraints, including scarce labeled data, expert-centered interpretation, long-tail class distributions, and strong sensitivity to image-acquisition conditions. These constraints influence how deep-learning methods should be selected, trained, evaluated, and interpreted [Fiorucci et al., 2020] [Colace et al., 2025].

This chapter connects the broader conceptual background to the methodological design of the thesis. It reviews deep-learning directions that directly affect cultural-heritage recognition and identifies the approaches most relevant to ancient Egyptian statue classification. The purpose is therefore not merely to catalogue previous publications, but to justify methodological choices related to transfer learning, data augmentation, multimodal extensions, explainability, and cautious evaluation [Fiorucci et al., 2020] [Paolanti et al., 2026].

Within cultural heritage, the transition toward deep learning is evident in applications involving monuments, sculptures, inscriptions, museum objects, archaeological artifacts, and restoration-related tasks. In these contexts, learned visual representations can capture complex patterns that are difficult to specify through manually designed features alone [Fiorucci et al., 2020] [Winterbottom et al., 2022].

The studies reviewed in this chapter address several related but distinct computer-vision tasks. Image classification assigns a category label to an entire input image, whereas object

detection identifies both the category and spatial location of each target object, commonly using bounding boxes. Image segmentation provides a finer spatial output by assigning labels at the pixel level. Semantic segmentation groups pixels according to their object category, while instance segmentation additionally separates individual objects belonging to the same category. Instance recognition addresses a different objective: it attempts to identify the exact physical artifact, monument, or landmark depicted in an image rather than assigning it only to a broader semantic category. Instance recognition should therefore not be confused with instance segmentation, which separates individual object regions at the pixel level but does not necessarily determine the exact historical identity of the depicted artifact.

These task formulations require different annotation structures, model outputs, and evaluation measures. Classification commonly uses accuracy, precision, recall, and F1-score. Detection additionally evaluates localization using measures such as Intersection over Union (IoU) and mean average precision (mAP). Segmentation typically relies on pixel-level measures such as IoU, the Dice coefficient, and pixel accuracy. Instance-recognition studies may instead use identification accuracy or retrieval-oriented measures, depending on whether the objective is direct identification or similarity-based retrieval. Consequently, the studies reviewed in this chapter should be interpreted according to their task formulation rather than compared solely through their reported numerical performance. The primary contribution of the present thesis is category-level image classification, while object detection is examined only as an exploratory baseline. A consolidated comparison of representative studies is provided later in Table 3.1, after the narrative review and comparative discussion.

## **3.2 CNN and Transfer Learning**

Transfer learning remains especially attractive for heritage imagery because specialist datasets are usually too small to justify training very deep models from scratch. Pretrained CNN backbones import generic visual priors such as edges, textures, and shape hierarchies, which can then be adapted to the target domain through fine-tuning. Studies on cultural-heritage building classification, museum-object analysis, and artifact recognition repeatedly report that transfer learning stabilizes optimization and improves final accuracy when

compared with naïve training on limited data [Belhi et al., 2021] [Winterbottom et al., 2022] [Ottoni and Ottoni, 2025].

Transfer learning also supports the comparison strategy used in this thesis. Rather than asking which architecture is universally best, the experiment asks which pretrained family adapts most effectively to this statue dataset under controlled training conditions. This is the basis for benchmarking EfficientNetV2B0, ResNet50V2, MobileNetV3Small, and VGG19 under a unified protocol [Belhi et al., 2021] [Ottoni and Ottoni, 2025].

CNN backbones pre-trained on large-scale datasets can be fine-tuned for heritage imagery, improving performance when labeled data is limited. Transfer learning is particularly useful when classes share low-level visual statistics with natural images but require domain-specific fine-grained discrimination [Belhi et al., 2021] [Winterbottom et al., 2022].

### 3.3 Generative Models and Data Augmentation

In cultural heritage research, data augmentation is not only a preprocessing detail; it is often a practical response to dataset sparsity, class imbalance, and narrow capture conditions. Geometric augmentations such as rotation, translation, and zoom can emulate moderate real-world variation, whereas photometric changes can reduce sensitivity to illumination bias. Ottoni and Ottoni [2025] explicitly analyze combinations of augmentations for heritage-building classification and show that augmentation strategy meaningfully affects generalization [Shorten and Khoshgoftaar, 2019] [Ottoni and Ottoni, 2025].

Generative models add a second line of inquiry. Rather than only perturbing existing images, they attempt to synthesize new samples or fill gaps in underrepresented classes. Recent museum-artifact work by Lu et al. [2026] illustrates how generative augmentation can strengthen multimodal classification under data scarcity. Even when a thesis does not directly train GANs or diffusion models, understanding this literature is valuable because it clarifies where simple augmentation may be sufficient and where richer synthesis becomes attractive.

GAN-based augmentation has been explored to address dataset scarcity and to improve

robustness to partial views and cropping. These approaches can be beneficial when original image acquisition is difficult or when datasets are highly imbalanced [Shorten and Khoshgoftaar, 2019] [Lu et al., 2026].

In the construction of EgyStatues-v1, generative models were not used to synthesize new statue images. Dataset-building augmentation was instead limited to conventional offline geometric transformations applied to selected underrepresented data using the TensorFlow/Keras ImageDataGenerator. The applied transformations included rotation, horizontal and vertical shifts, shear, zoom, and horizontal flipping. In the retained augmentation run, twenty transformed samples were generated from each of forty-two source images belonging to an underrepresented class, producing 840 additional images. These transformed images were incorporated into the dataset before the final train-validation-test partition was created. This approach increased viewpoint and scale variation while avoiding the introduction of synthetic statue content whose historical or iconographic validity had not been independently verified. Generative adversarial networks and diffusion-based synthesis are therefore discussed in this chapter as relevant future directions rather than as components of the released EgyStatues-v1 benchmark. The detailed augmentation configuration and its associated split-integrity limitation are reported in Section 4.2.2.3.

### **3.4 3D and Multimodal Learning**

Heritage objects are documented not only through images but also through text, provenance, material descriptors, dimensions, and sometimes 3D scans. Multimodal approaches, therefore, provide a natural extension of vision-only classification. Rei et al. [2023] demonstrate that combining image, text, and tabular information improves metadata assignment for cultural-heritage artifacts, while Lu et al. [2026] use cross-modal attention to couple image and textual cues for museum artifact classification.

Three-dimensional learning is equally relevant because statues are intrinsically volumetric objects whose appearance varies strongly across viewpoints. Although the present thesis uses 2D photographs, future systems could benefit from photogrammetry, point clouds, or multi-view learning when richer acquisition is available. The literature on point-cloud

segmentation, 3D reconstruction, and geometry-aware heritage analysis suggests that a purely 2D benchmark is an important starting point, but not the endpoint of the field [Rei et al., 2023] [Xu et al., 2025].

Because statues are inherently three-dimensional objects, 3D-aware approaches and multimodal fusion (e.g., image + textual metadata) represent promising directions to reduce ambiguity among visually similar categories [Rei et al., 2023] [Xu et al., 2025].

### **3.5 Explainable AI for Trustworthy Heritage Systems**

Interpretability has special weight in heritage domains because model outputs may influence scholarly search, educational content, or conservation workflows. Saliency maps, Grad-CAM visualizations, and concept-level explanations cannot replace expert judgment, but they can reveal whether a classifier attends to plausible object regions or to spurious background cues. This matters in datasets assembled from web imagery, where museums, display cases, and photographic framing may unintentionally correlate with labels [Paolanti et al., 2026] [Gokak et al., 2025].

Recent work on trustworthy AI in cultural heritage argues that fairness, explainability, and bias mitigation should be integrated into the design process rather than treated as post-deployment additions. Paolanti et al. [2026] frame this explicitly as a methodological requirement for the sector. This perspective supports the thesis decision to include qualitative predictions and critical error analysis instead of reporting aggregate metrics alone.

Interpretability is important in academic heritage contexts where experts need to understand the visual evidence supporting a prediction. Explainable AI techniques, such as saliency maps and concept-based explanations, can help bridge the gap between model outputs and domain reasoning [Selvaraju et al., 2017] [Paolanti et al., 2026]. The studies supporting these methodological directions are discussed in Sections 3.6 and 3.7 and are consolidated comparatively in Section 3.8.

### 3.6 Detailed Review of Recent Studies

To position the thesis within contemporary research, this section synthesizes recent work related to cultural heritage recognition, landmark recognition, museum-object classification, and detection in real-world scenes. The aim is selective rather than exhaustive: the reviewed studies are used to extract methodological patterns that influence the present benchmark, including transfer learning, augmentation, lightweight deployment, transformer-based representations, and the distinction between classification and detection.

A useful way to read the recent literature is to group it by problem structure rather than by application label. One cluster focuses on category recognition, where the goal is to assign an object or scene to a defined class. A second cluster addresses instance-level recognition and retrieval, where the system must distinguish one exact object from many visually related candidates. A third cluster centers on detection or segmentation, where localization matters as much as class prediction. Heritage applications can belong to any of these clusters, and methodological choices become clearer when the problem formulation is made explicit.

The present thesis sits primarily in the category-recognition cluster, but it borrows lessons from the other two. Instance-recognition studies highlight the difficulty of fine-grained visual distinctions under real-world image variation. Detection studies clarify when localization is useful and when it adds unnecessary complexity. Reading the literature through this task structure prevents the thesis from importing methods simply because they are popular; instead, the chosen model family is evaluated against the actual properties of the statue-recognition problem.

#### 3.6.1 Heritage building and architectural element classification

Several recent studies focus on classifying heritage buildings or their architectural components using transfer learning. These works often face constraints similar to those in museum artifact recognition: limited labeled data, strong appearance variation due to illumination and viewpoint, and the need for practical deployment. Ottoni and Ottoni [2025] report that the choice and combination of augmentation operations can produce substantial gains and propose a dedicated dataset for colonial architecture. For the present benchmark, this

supports treating augmentation as an explicit experimental variable rather than as a fixed preprocessing default.

### **3.6.2 Detection-based heritage documentation**

Object detection has become a standard tool for documenting heritage sites, identifying components (e.g., windows, doors, decorative elements), and supporting condition monitoring (crack detection, damage localization). Du and Wang [2024] propose Bi-YOLO alongside a dataset tailored to heritage building components. Gao et al. [2024] adapt YOLOv8 for garden element detection under occlusions and clutter. These studies motivate the inclusion of YOLO baselines while also helping explain why detection may underperform when images predominantly contain a single centered object.

### **3.6.3 Museum artifact classification and multimodal fusion**

Museum collections often provide more than images: accession metadata, material labels, and textual descriptions. Lu et al. [2026] propose a cross-modal attention fusion approach with generative augmentation to address limited labeled data. In settings where such metadata is available, multimodal fusion can improve robustness and reduce confusions between visually similar classes. The present thesis focuses on the image-only case to ensure reproducibility from public web images, but the dataset design can be extended with object descriptions, acquisition metadata, and curator-reviewed textual attributes.

### **3.6.4 Landmark recognition and instance-level retrieval**

Landmark recognition is closely related to statue classification because it is a fine-grained recognition problem under uncontrolled capture conditions. GLDv2 [Weyand et al., 2020] formalizes large-scale instance recognition and retrieval, encouraging evaluation under domain shift. The heritage field can borrow practices from landmark recognition, including strict train/test splits, hard negative analysis, and evaluation that reflects real usage. These practices inform the thesis emphasis on fixed splits and on distinguishing category recognition from instance retrieval.

### **3.6.5 Transformers and large vision models in heritage**

Transformer-based models (ViT and related architectures) are increasingly used for heritage recognition. Xu et al. [2025] present ViT-HVE, showing that transformers can capture global context relevant for heritage value analysis. Although CNN transfer learning remains highly competitive for many classification tasks, transformer approaches are a credible direction for future work, especially when large-scale pretraining or multimodal alignment (image + text) is available. This makes transformer evaluation a natural extension once the CNN baseline and dataset protocol are stable.

### **3.6.6 Surveys and bibliometric mappings**

Surveys help identify dominant methods and open challenges. Ali and Zhang [2024] provide a comprehensive review of YOLO variants up to YOLO11, clarifying the trade-off between speed and accuracy that is central to deployable heritage applications. Colace et al. [2025] discuss broader AI challenges for heritage, including data governance and sustainability considerations. Bibliometric mapping studies also highlight emerging clusters such as multimodal learning, self-supervised pretraining, and explainable AI. Together, these sources support the thesis's focus on governance, transparency, and carefully delimited claims.

## **3.7 Additional Relevant Studies**

This section discusses additional studies that contribute to the methodological and conceptual design of the thesis. The selected works cover archaeological conservation, cultural-heritage protection, object detection, museum-artifact classification, multimodal learning, explainable artificial intelligence, and deployment-oriented evaluation. Together, they provide broader context for the dataset construction, model selection, experimental evaluation, and future research directions adopted in this thesis.

Casillo et al. [2025] frame the application of artificial intelligence in archaeological site conservation as an interdisciplinary workflow in which computer vision, data quality, annotation standards, and stakeholder collaboration must be considered together. Their analysis indicates that the value of an artificial intelligence system depends not only on its

predictive performance but also on the transparency of its data and methodological decisions. This perspective supports the present thesis by emphasizing the importance of reproducible dataset construction, explicit reporting of limitations, and clearly defined boundaries for the use of classification outputs in cultural heritage applications.

Similarly, Colace et al. [2025] provide a broad overview of emerging artificial intelligence challenges in cultural heritage protection, including predictive maintenance, Internet of Things-based monitoring, and explainable, human-centered decision support. Their review identifies practical limitations such as the scarcity of labeled data, domain shift, and ethical concerns associated with automated decision-making. These observations support the thesis emphasis on balancing classification accuracy with computational practicality, transparent evaluation, and the continued involvement of heritage professionals in interpreting model outputs.

From a detection-oriented perspective, Gao et al. [2024] propose an improved YOLOv8-based detector for recognizing elements of traditional private gardens in Jiangnan. Their approach enhances feature fusion and detection heads, resulting in improved mean average precision. In addition to the architectural contribution, the study describes a domain-specific annotation scheme, dataset construction procedures, and augmentation strategies intended to improve robustness across variations in viewpoint and illumination. This work is conceptually relevant to the exploratory YOLO experiments in the present thesis because it demonstrates how YOLO-family models can support heritage documentation when object localization is required.

Lu et al. [2026] address museum artifact classification through cross-modal attention fusion and generative data augmentation. Their approach combines visual and contextual information to improve recognition under class imbalance and intra-class variability. The reported results suggest that multimodal representations can be especially valuable when artifacts share similar visual shapes, materials, or decorative characteristics. Although the present thesis focuses primarily on vision-based classification, this study provides a strong motivation for future extensions that combine statue images with textual descriptions, historical metadata, provenance information, or museum catalog records.

Rei et al. [2023] also investigate multimodal cultural heritage classification by combining a ResNet-based image model, a multilingual text model, and structured tabular features through late fusion. Their study emphasizes the importance of taxonomy-aware dataset design and demonstrates that combining multiple data modalities can improve missing-property prediction compared with unimodal approaches. These findings support the thesis discussion of scalability and suggest that the proposed statue recognition framework could eventually be extended from visual category prediction to richer museum cataloging and metadata-assignment tasks.

Transformer-based approaches are represented by Xu et al. [2025], who propose the ViT-HVE framework for cultural heritage recognition and value evaluation. The study argues that attention mechanisms can capture global visual context and long-range dependencies that may not be represented as effectively by conventional convolutional operations. This is particularly relevant when images contain complex visual compositions or when discriminative information is distributed across distant regions. For the present thesis, the study supports the inclusion of Vision Transformer architectures as an important future extension beyond the current CNN-based transfer-learning benchmark.

Chen et al. [2025] examine deep learning-based pathology detection and analysis in historic masonry buildings. Their work highlights variability in image acquisition conditions and stresses the importance of aligning model predictions with practical conservation workflows. The study is relevant to the present thesis because it demonstrates that model evaluation should extend beyond a single overall accuracy value. Class-specific behavior, qualitative inspection, error analysis, and the practical interpretation of predictions are also necessary when artificial intelligence systems are intended to support heritage professionals.

Boutet and Hallot [2025] propose a practical framework for applying YOLOv5 to crack detection in built heritage and investigate the effect of ground sampling distance on detection performance. Their analysis demonstrates that image acquisition conditions can substantially influence the reliability of models deployed in real inspection environments. The emphasis placed on acquisition settings, deployment constraints, and qualitative evaluation is consistent with the present thesis approach of documenting the computational environment and

discussing the practical limitations of the exploratory detection baselines.

Du and Wang [2024] introduce Bi-YOLO together with a dedicated dataset for detecting architectural components in heritage buildings. Their method modifies the feature-pyramid design to improve the localization of structured architectural elements and supports automated indexing and documentation. This work complements the YOLO-related experiments in the thesis and reinforces the broader conclusion that detection models are particularly valuable when the task requires the location of multiple objects or architectural components within a scene. However, this differs from the main task of the present study, in which most images contain a dominant statue and the primary objective is category classification.

Gokak et al. [2025] present a lightweight MobileNetV3-based approach for classifying defects in heritage building images and use Grad-CAM heatmaps to provide visual explanations for model decisions. Their work demonstrates that compact neural networks can offer a useful balance between predictive performance, computational efficiency, and interpretability. This finding is consistent with the inclusion of MobileNetV3Small in the present thesis as an efficient transfer-learning baseline and supports the broader argument that deployment suitability should be considered alongside classification accuracy.

Ali and Zhang [2024] provide a comprehensive review of the YOLO framework, covering its architectural evolution, recent versions, performance trends, and deployment considerations. Their review provides theoretical and technical context for the exploratory use of YOLOv11 and YOLOv12 in the present thesis. It also explains why one-stage object detectors remain attractive for real-time cultural heritage documentation, particularly when rapid inference and localization are required.

Murat and Kiran [2025] similarly provide a systematic review of YOLO architectures from YOLOv1 to YOLOv11, with particular emphasis on recent innovations and limitations in existing comparative studies. Their analysis highlights the trade-off between detection accuracy and computational speed across different YOLO variants. This is relevant to the experimental discussion in the thesis because it reinforces the need to document deployment-related variables such as inference latency, image resolution, batch size, and computational

hardware when assessing detection models.

Hasan et al. [2025] discuss the continued evolution of YOLO architectures up to YOLOv12 and summarize the major design changes introduced across different versions. Because this work was published as a preprint, it is treated in the thesis as supplementary technical context rather than as a primary source of established evidence. Its role is limited to positioning the exploratory YOLOv12 experiment within the broader development of the YOLO family, while the conclusions of the thesis remain grounded in the reported experimental results and more established peer-reviewed literature.

Beyond visual recognition, Dragomir and Dragomir [2025] propose a hybrid neuro-fuzzy system for the proactive maintenance of heritage sites. Their work emphasizes interpretability, risk assessment, and decision support in conservation planning. Although the study is not centered exclusively on computer vision, it illustrates a broader application context in which image-based recognition systems may provide input to maintenance planning and heritage risk-management platforms. This perspective supports the view that statue recognition can serve as one component within a larger human-supervised heritage information system.

Wen et al. [2024] investigate museum artifact recognition in an interactive setting and propose an adaptive CNN architecture with adaptive pooling to improve robustness. Their study demonstrates how automated artifact recognition can be integrated into educational, multimedia, and visitor-support applications within museums. This practical setting supports the motivation of the present thesis, particularly its focus on digital documentation, museum cataloging, and heritage-oriented image retrieval.

Recent semantic-segmentation studies in heritage monitoring further demonstrate that cracks, deterioration patterns, and surface defects in monuments may differ substantially from similar patterns found in conventional infrastructure. This difference creates domain-shift challenges and increases the need for specialized annotated datasets. Such work suggests that future versions of the proposed dataset could include pixel-level annotations, bounding boxes, or weakly supervised localization when the research objective extends beyond statue category

classification to condition assessment or detailed surface monitoring [Chen et al., 2025; Boutet and Hallot, 2025].

Large vision models and foundation models are also becoming increasingly relevant to cultural heritage content analysis. Their scalability, transfer-learning capabilities, and potential for self-supervised and multimodal learning may reduce dependence on large manually labeled datasets. These models could support future extensions of the present work to larger and more diverse artifact collections, particularly when combined with textual museum records or cross-collection retrieval mechanisms [Radford et al., 2021; Dosovitskiy et al., 2021; Xu et al., 2025].

Wisniamongkol et al. [2024] study cultural heritage image classification under class-imbalance conditions and discuss training practices that can stabilize learning when some classes contain substantially fewer examples than others. Their findings support the dataset-construction decisions adopted in the thesis, including the use of reviewed augmentation for underrepresented categories and the preservation of class representation across the training, validation, and test partitions. The study also reinforces the importance of reporting class-wise performance rather than relying exclusively on aggregate accuracy.

Fiorucci et al. [2020] provide a broad survey of machine learning applications in cultural heritage, covering classification, retrieval, reconstruction, restoration, and related analytical tasks. Although the study predates several recent deep learning developments, it remains an important reference because it identifies persistent challenges such as data scarcity, model interpretability, heterogeneous image sources, and domain shift. These challenges remain directly relevant to the construction and evaluation of the statue-recognition benchmark developed in this thesis.

Finally, Vijayakumar and Vairavasundaram [2024] review YOLO-based object detection models, their recent architectural improvements, major application areas, and commonly used evaluation practices. Their analysis complements the review provided by Ali and Zhang [2024] and offers additional context for interpreting detection-model performance and deployment requirements. Together, these reviews support the thesis decision to treat YOLO models as exploratory detection baselines rather than as directly equivalent alternatives

to the main image-classification models.

Taken together, the reviewed studies clarify the specific research gap addressed by this thesis. Existing cultural heritage research demonstrates that deep learning can support classification, detection, segmentation, monitoring, and multimodal cataloging across a variety of heritage contexts. However, fewer studies provide a focused taxonomy of ancient Egyptian statues, a documented web-to-benchmark curation process, a controlled comparison across multiple CNN-based transfer-learning models, and an explicit discussion of why object detection may not be the most suitable initial formulation for predominantly single-object images. The present thesis addresses this gap by organizing ancient Egyptian statues into twenty-five interpretable categories, treating dataset construction as a central research contribution, and interpreting experimental performance within the practical contexts of museum cataloging, digital archiving, image retrieval, and expert-assisted heritage documentation.

## **3.8 Comparative Synthesis of the Literature and Methodological Lessons**

### **3.8.1 Recurring Methodological Limitations**

At the same time, a number of weaknesses persist. Many datasets remain small, narrow, or geographically restricted. Some studies emphasize impressive metrics without documenting class overlap, ambiguous annotations, or the effect of source bias. Others focus on technical novelty while leaving deployment questions unresolved: how would the system behave with an unseen category, an unusually cropped image, or a low-quality mobile capture from a museum visitor? In addition, explainability is still unevenly integrated. Even when accuracy is strong, heritage professionals may hesitate to rely on a system that cannot provide understandable evidence for its suggestions. These gaps justify continued attention to benchmark documentation, class definition, and human-in-the-loop validation [Mishra and Lourenço, 2024] [Paolanti et al., 2026].

### **3.8.2 Recurring Strengths in Recent Cultural-Heritage Studies**

Across the stronger studies, several strengths recur. First, successful papers define the

heritage asset carefully instead of using broad labels that mix heterogeneous objects. Second, they connect model choice to a concrete use case such as inventory support, visual inspection, monument understanding, or artifact retrieval. Third, they document their training settings sufficiently well for replication. Fourth, they recognize that visual AI in heritage should support, not replace, expert judgment. These strengths are visible in studies of antiquities classification, heritage-building categorization, multimodal museum-artifact analysis, and damage-detection workflows [Winterbottom et al., 2022] [Siountri and Anagnostopoulos, 2023] [Mnasri and D'Andrea, 2025] [Paolanti et al., 2026].

### 3.8.3 Implications for the Present Thesis

For this thesis, the most useful lesson is that benchmark quality and comparability are themselves contributions. Several heritage studies report strong performance but are difficult to compare directly because they use different data splits, varying preprocessing pipelines, inconsistent augmentation, or limited discussion of class ambiguity. By fixing the dataset split, input scale, task definition, and evaluation logic across several CNN backbones while openly acknowledging model-specific preprocessing differences, the present work addresses a practical gap in the literature: heritage practitioners often need a dependable baseline and transparent methodology before they need a highly customized architecture [Winterbottom et al., 2022] [Ottoni and Ottoni, 2025].

The reviewed studies also suggest that there is no single dominant architecture for all heritage problems. CNN transfer learning remains highly competitive when datasets are moderate in size and the task is visually coherent. Detection-oriented models are advantageous when the image contains multiple relevant regions or when automatic localization is essential. Multimodal fusion becomes attractive when descriptive text, metadata, or provenance information is available. Transformer-based approaches are promising, especially when long-range contextual relations matter, but their benefits are often moderated by data volume, computational cost, and the need for careful regularization. Consequently, model choice should follow task structure, available annotations, and deployment constraints rather than current hype cycles alone [Fiorucci et al., 2020] [Colace et al., 2025].

A cross-reading of the literature reveals a clear transition in cultural-heritage artificial intelligence. Early work often focused on demonstrating that pretrained CNNs could be transferred successfully to heritage datasets. More recent studies address broader questions concerning heterogeneous sources, multimodal fusion, explainability, bias awareness, and benchmark design beyond a single controlled setting. This shift positions the present work beyond merely comparing several backbones and toward constructing a documented classification benchmark that is interpretable, reproducible, and relevant to digital cataloging [Fiorucci et al., 2020] [Colace et al., 2025]. Table 3.1 consolidates representative studies after their methodological contexts have been discussed. The studies are arranged chronologically by publication year, with studies from the same year ordered alphabetically by the surname of the first author.

Table 3.1: Consolidated comparison of representative cultural-heritage deep-learning studies reviewed in this chapter.

| Study                   | Heritage asset                 | Task                               | Data setting                      | Modeling direction          | Main lesson                                                                                        | Relevance here                                                                          |
|-------------------------|--------------------------------|------------------------------------|-----------------------------------|-----------------------------|----------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| [Fiorucci et al., 2020] | Cross-domain cultural heritage | Survey                             | Multiple datasets                 | Methodological survey       | Data quality, domain specificity, and interpretability remain central concerns.                    | Supports benchmark-centered framing and cautious evaluation.                            |
| [Weyand et al., 2020]   | Landmarks                      | Instance recognition and retrieval | Approximately 5 million images    | CNN and retrieval pipelines | Large-scale instance recognition requires strict evaluation under domain shift and hard negatives. | Clarifies the distinction between category classification and exact-instance retrieval. |
| [Belhi et al., 2021]    | Cultural-heritage images       | Classification                     | Moderate-sized curated image sets | Pretrained CNNs             | Transfer learning provides a strong baseline when labeled heritage data are limited.               | Motivates CNN-based transfer-learning benchmarking.                                     |

| Study                                | Heritage asset              | Task                       | Data setting                          | Modeling direction                    | Main lesson                                                                                               | Relevance here                                                                  |
|--------------------------------------|-----------------------------|----------------------------|---------------------------------------|---------------------------------------|-----------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| [Winterbottom et al., 2022]          | Museum antiquities          | Instance classification    | Heterogeneous object photographs      | Deep image classification             | Heritage objects require robust fine-grained representations under variable capture conditions.           | Provides a close artifact-recognition precedent.                                |
| [Rei et al., 2023]                   | Museum metadata assignment  | Multimodal classification  | Image, text, and tabular metadata     | Late-fusion multimodal learning       | Combining visual and contextual modalities can improve metadata assignment.                               | Motivates future multimodal statue cataloging.                                  |
| [Siountri and Anagnostopoulos, 2023] | Heritage buildings          | Classification             | Street-view-style imagery             | Deep-learning and YOLO-based pipeline | Class overlap and urban variability require domain-specific labels and evaluation.                        | Illustrates labeling and domain-variation challenges.                           |
| [Gao et al., 2024]                   | Traditional private gardens | Object detection           | Object-level annotated scenes         | Optimized YOLOv8                      | Detection is effective when heritage elements must be localized in complex backgrounds.                   | Explains why detection is an exploratory baseline rather than the primary task. |
| [Mishra and Lourenço, 2024]          | Built heritage              | Review                     | Inspection-focused literature         | AI-assisted visual inspection         | Deployment requires interpretability, risk awareness, and realistic evaluation.                           | Informs the validity and deployment discussion.                                 |
| [Chen et al., 2025]                  | Historic masonry buildings  | Detection and segmentation | 895 images across 5 pathology classes | YOLOv8l-seg                           | Heritage-specific pathology classes can be localized, but acquisition variability affects interpretation. | Highlights the different annotation and evaluation needs of localization tasks. |

| Study                       | Heritage asset                      | Task                          | Data setting                           | Modeling direction                                | Main lesson                                                                          | Relevance here                                                                                                 |
|-----------------------------|-------------------------------------|-------------------------------|----------------------------------------|---------------------------------------------------|--------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
| [Colace et al., 2025]       | Cross-domain cultural heritage      | Review                        | Recent cultural-heritage AI literature | AI-applications overview                          | The field is moving toward broader integration, governance, and evaluation.          | Supports the research-gap analysis.                                                                            |
| [Mnasri and D'Andrea, 2025] | Archaeological artifacts            | Detection and classification  | Inventory-oriented dataset             | Deep and transfer learning                        | Pipeline choice should follow the intended inventory and localization workflow.      | Provides the closest mixed-task inventory comparison.                                                          |
| [Ottoni and Ottoni, 2025]   | Historic buildings                  | Classification                | 6,449 images across 5 classes          | Transfer learning with augmentation               | Augmentation strategy interacts strongly with dataset design and generalization.     | Supports the preprocessing and augmentation rationale.                                                         |
| [Xu et al., 2025]           | Cultural-heritage value recognition | Transformer-based recognition | Specialized heritage imagery           | ViT-HVE framework                                 | Transformer representations are promising but remain data- and context-dependent.    | Places the retained CNN benchmark in a wider modeling context.                                                 |
| [Lu et al., 2026]           | Museum artifacts                    | Multimodal classification     | Image and textual museum data          | Cross-modal attention and generative augmentation | Multimodal fusion and augmentation can improve recognition under data scarcity.      | Motivates future multimodal and generative extensions; these methods were not used to construct EgyStatues-v1. |
| [Paolanti et al., 2026]     | Cross-domain cultural heritage      | Trustworthy-AI framework      | Mixed case-oriented evidence           | Explainability and bias mitigation                | Transparency, bias awareness, and expert oversight are central to heritage adoption. | Supports human-in-the-loop interpretation and responsible deployment boundaries.                               |

Table 3.1 is a qualitative synthesis rather than a numerical ranking. The studies address different tasks, annotation structures, datasets, and evaluation measures; their reported performance values should therefore not be compared as though they were produced under a common benchmark.

### 3.9 Research Gaps and Motivation for the Present Thesis

The cross-study limitations identified in Section 3.8 are translated here into thesis-specific research gaps. Table 3.2 maps each gap to the corresponding methodological response of the present work.

Table 3.2: Main research gaps identified in the cultural-heritage literature and the corresponding contribution of the present thesis.

| Observed Gap in the Literature                                                      | How the Present Thesis Responds                                                        |
|-------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------|
| Many CH datasets are small, private, or weakly documented.                          | Introduces a comparatively large, documented benchmark focused on statue categories.   |
| Category labels are often broad or inconsistent.                                    | Uses a statue-centered taxonomy and manual curation to improve label coherence.        |
| Studies often optimize one architecture without controlled comparison.              | Benchmarks multiple pretrained CNN families and detector baselines under one protocol. |
| Evaluation frequently emphasizes overall accuracy only.                             | Adds qualitative examples, per-class discussion, and detector metrics where relevant.  |
| Multimodal and trustworthy-AI directions are discussed more often than implemented. | Positions them explicitly as justified future extensions and design considerations.    |

These gaps are addressed through three connected components of the thesis. First, the literature-review component synthesizes the broader field and clarifies why dataset quality, transfer learning, explainability, and trustworthy deployment are important in cultural-heritage applications. Second, the dataset-construction component introduces EgyStatues-v1 as a statue-specific benchmark with an explicit class structure and documented curation process.

Third, the experimental component evaluates multiple CNN-based transfer-learning classifiers under a common benchmark and interprets their behavior using class-wise analysis rather than relying solely on aggregate rankings. Together, these components provide the scientific and methodological foundation of the thesis.

A further gap concerns evaluation framing. Some studies in the broader literature overgeneralize from narrow datasets or present results without clearly separating formal claims from exploratory side experiments. The present thesis therefore states its scope boundaries explicitly: the retained core benchmark is CNN-centered, the YOLO runs are discussed as detection-oriented side experiments related to task formulation, and transformer-based or multimodal extensions are positioned as justified next steps rather than unsupported claims [Mishra and Lourenço, 2024] [Paolanti et al., 2026].

### **3.10 Summary**

This chapter reviewed the principal deep-learning directions relevant to cultural-heritage recognition, including transfer learning, data augmentation, multimodal learning, generative methods, and explainable artificial intelligence. It also synthesized representative studies, identified recurring methodological limitations, and clarified the research gaps addressed by the present thesis. The next chapter presents the construction, curation, split organization, and evaluation measures of EgyStatues-v1, thereby establishing the data and metric foundation for the proposed framework in Chapter 5.

## **Chapter 4**

### **Dataset Construction and Performance Measures**

#### **4.1 Introduction**

Benchmark datasets and clear performance measures are essential for reproducible computer-vision research. In cultural-heritage applications, the scarcity of standardized datasets often limits direct comparison across approaches and makes it difficult to judge whether apparent progress comes from better models, better data, or easier evaluation settings. This chapter, therefore, defines the dataset and metric foundation of the thesis: it introduces the proposed EgyStatues-v1 benchmark, explains the curation logic behind it, and summarizes the evaluation measures used in the later experiments [Fiorucci et al., 2020] [Lamas et al., 2021].

In heritage informatics, the dataset is often as important as the model itself. When class definitions are unstable, provenance is unclear, duplicates remain uncontrolled, or train/test partitions do not reflect the intended task, reported performance can be misleading. For that reason, the present chapter treats dataset construction not as a preliminary technical step but as a core scientific contribution. This emphasis is consistent with the literature review, which identifies benchmark scarcity and deployment realism as recurring weaknesses, and with the dataset-construction component of the thesis, which formalizes EgyStatues-v1 as a reusable benchmark resource [Mishra and Lourenço, 2024] [Paolanti et al., 2026].

#### **4.2 Dataset Construction**

##### **4.2.1 Related Cultural Heritage Datasets (Brief Overview)**

Prior work in cultural heritage has released datasets targeting related tasks such as monument recognition, hieroglyph classification, and artifact material identification. However, statue-specific datasets are typically small, heterogeneous, or not publicly standardized, which

motivates the development of a dedicated benchmark for ancient Egyptian statues [Fiorucci et al., 2020] [Lamas et al., 2021].

Relevant precedents include landmark-recognition benchmarks, museum-object datasets, and archaeological image collections. These resources are useful for methodological inspiration, yet they do not directly solve the problem addressed here. Landmark datasets emphasize place-level retrieval, while museum datasets often mix heterogeneous objects, materials, and metadata conditions. A dedicated statue-focused benchmark is therefore justified when the target is fine-grained recognition within a coherent artifact family [Weyand et al., 2020] [Winterbottom et al., 2022].

To clarify the relationship between existing cultural-heritage datasets and the benchmark proposed in this thesis, Table 4.1 compares representative datasets covering Egyptian monuments, museum artifacts, ancient Egyptian hieroglyphs, Egyptian shabtis, and ancient Egyptian statue categories.

Table 4.1: Comparison of representative cultural-heritage datasets related to EgyStatues-v1.

| Dataset / Study                   | Heritage focus                                                                  | Main task                                     | Dataset scale                                         | Main difference from EgyStatues-v1                                                                                                                    |
|-----------------------------------|---------------------------------------------------------------------------------|-----------------------------------------------|-------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
| EGYPT-v1<br>[Hassan et al., 2023] | Ancient Egyptian monuments, landmarks, heritage sites, and mixed artifact types | Image classification and monument recognition | More than 7,000 images across approximately 40 labels | It covers a broad mixture of pyramids, temples, statues, mummies, palaces, and heritage sites rather than a controlled taxonomy of statue categories. |

| Dataset / Study                                              | Heritage focus                                                | Main task                                                       | Dataset scale                                                   | Main difference from EgyStatues-v1                                                                                                                                   |
|--------------------------------------------------------------|---------------------------------------------------------------|-----------------------------------------------------------------|-----------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Oriental Museum Artefact Dataset [Winterbottom et al., 2022] | Individual museum artifacts from several cultural collections | Artefact instance classification                                | 24,502 images representing 4,332 individual object instances    | It aims to recognize the exact identity of previously observed objects, whereas EgyStatues-v1 performs category-level classification of statue types.                |
| Ancient Egyptian Hieroglyph Dataset [Barucci et al., 2021]   | Individual ancient Egyptian hieroglyphic signs                | Hieroglyph image classification                                 | 4,309 images distributed across 40 classes                      | It focuses on symbolic written signs rather than complete three-dimensional statue objects and therefore addresses a substantially different visual task.            |
| Egyptian Shabti Dataset [Duque Domingo et al., 2020]         | Egyptian funerary shabtis and their hieroglyphic names        | Detection, owner identification, and dynasty/period recognition | 815 training images, 354 test images, and 158 shabti categories | It is highly specialized in identifying individual shabti owners and inscriptions, while EgyStatues-v1 covers a broader range of ancient Egyptian statue categories. |

| Dataset / Study | Heritage focus                                     | Main task                                        | Dataset scale                       | Main difference from EgyStatues-v1                                                                                                                                                                                                                                       |
|-----------------|----------------------------------------------------|--------------------------------------------------|-------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| EgyStatues-v1   | Ancient Egyptian statues and statue-like artifacts | Fine-grained category-level image classification | 41,161 RGB images across 25 classes | It provides a large statue-specific taxonomy, manual label verification, duplicate control, documented curation, and a fixed train-validation-test organization for comparative benchmarking, while source-family independence across the partitions was not guaranteed. |

The comparison shows that the existing datasets address important but distinct cultural-heritage recognition problems. Monument datasets generally combine heterogeneous architectural sites, landmarks, statues, and other heritage assets under a broad recognition setting. Artefact datasets often focus on instance-level identification, where the objective is to recognize a specific museum object rather than assign it to a broader semantic category. Hieroglyph datasets address the classification of symbolic written signs, while shabti datasets concentrate on a highly specialized artifact type and may additionally incorporate owner, inscription, dynasty, or historical-period information. In contrast, EgyStatues-v1 is specifically designed for fine-grained category-level classification of ancient Egyptian statues. Among the representative datasets compared in Table 4.1, it provides the largest image collection and combines a controlled twenty-five-class taxonomy with manual curation, duplicate filtering, class-aware dataset construction, and a documented train-validation-test organization. However, because offline augmentation preceded the final split, the archived organization does not guarantee that every source image and all of its transformed derivatives remained within the same partition. These characteristics make it more directly suitable for benchmarking deep-learning models on statue-specific recognition rather than on broader monument, symbolic, or

exact-object identification tasks.

### 4.2.2 The Proposed EgyStatues-v1 Dataset

EgyStatues-v1 is a curated benchmark of 41,161 images distributed across 25 ancient Egyptian statue categories. It was assembled through class-guided retrieval from publicly accessible museum pages, digital-heritage repositories, and targeted web-image search, followed by automated filtering and manual verification. The class structure focuses on statue-specific categories rather than broad monument or museum-object groups, because the thesis addresses fine-grained single-object recognition rather than general heritage classification. After curation, the dataset was partitioned into 70% training, 15% validation, and 15% testing subsets so that each class remained represented across all splits. The same retained version is used throughout the experimental analysis.

The selection of the twenty-five statue categories followed a purposive, prominence-based strategy rather than a random sampling procedure. An initial candidate list was compiled from well-known and frequently documented ancient Egyptian statues based on their prominence in museum collections, heritage literature, and public digital archives. Candidate classes were subsequently filtered according to data-sufficiency criteria: categories for which insufficient or unreliable image evidence could be retrieved were excluded from the final taxonomy. This two-stage process—prominence-driven candidate identification followed by availability-based filtering—produced the 25 statue categories listed in Table 4.2.

Using public web imagery gives the dataset a degree of ecological validity: the collected images resemble the heterogeneous photographs that real systems may encounter in search engines, collection portals, digital catalogs, or educational interfaces. The same choice also introduces risks, including source bias from museum presentation styles, repeated publicity photographs, and the uneven online visibility of famous classes. For that reason, the curation protocol aimed to preserve breadth while enforcing class coherence, and the benchmark should be understood as a filtered research dataset rather than a raw web crawl [Mishra and Lourenço, 2024] [Wisnongkol et al., 2024].

#### 4.2.2.1 Data Acquisition via Automated Web Scraping

To construct EgyStatues-v1, a class-first image-acquisition workflow was implemented to collect candidate images for each of the twenty-five target statue categories. The majority of the candidate images were acquired through general web-image retrieval using DuckDuckGo, while a smaller supplementary portion was obtained from publicly accessible museum collection pages and heritage-related websites. The workflow was executed in Google Colab using the open-source `jmd_imagescraper` utility, and the downloaded images were stored in a separate folder corresponding to each class label.

The retrieval strategy used the English name of each target class as both the folder label and the search query. For example, the Block Statue class was collected using the following call: `duckduckgo_search(root, "Block Statue", "block statue", max_results=1000)`. The same procedure was repeated for the remaining classes by replacing the folder name and query with class-specific terms such as “Great Sphinx of Giza,” “Statue of Anubis,” “Nefertiti Bust,” and the other category names listed in Table 4.2. This approach ensured that each class was treated as an independent retrieval unit and simplified subsequent class-wise inspection and organization.

The `max_results` parameter was set to a high upper limit of 1,000 candidate results for each query to improve potential image coverage. However, this value represented only the maximum requested number of results and did not guarantee that 1,000 valid files would be downloaded for every class. The actual number varied according to search availability, inaccessible links, repeated results, and download failures. For example, the archived Block Statue retrieval run returned 384 downloaded files despite using `max_results=1000`.

No explicit pre-download restrictions were applied to image size, image type, layout, or color in the class-acquisition calls; therefore, the scraper’s default retrieval settings were used. This broad acquisition strategy was intended to capture heterogeneous photographs with variations in viewpoint, lighting, background, scale, and photographic quality. Images collected from museum and heritage pages were supplementary to the general web-scraping process rather than being selected according to a fixed provider priority.

The exact total number of raw images collected across all twenty-five classes before cleaning and augmentation was not retained in the archived acquisition records and therefore cannot be reconstructed reliably. In addition, the final benchmark size should not be interpreted as the original download count because the retained dataset of 41,161 images includes both curated source images and reviewed offline-augmented samples generated for selected underrepresented classes. This limitation is reported explicitly to avoid inferring an unsupported raw-image count from the final dataset size.

Source inclusion was based primarily on public accessibility rather than on a fixed hierarchy of websites. Candidate images from general web results, museum pages, and heritage-related websites could enter the initial collection pool. Final retention decisions were made at the image level during the subsequent cleaning and curation stage. A sample was retained only when it showed a sufficiently clear and recognizable statue, corresponded to the intended class, and provided enough visual information for image-level classification. Irrelevant search results, incorrectly labeled images, ambiguous objects, collages, thumbnails, duplicate or near-duplicate samples, and severely degraded images were removed as described in Section 4.2.2.2.

#### **4.2.2.2 Manual Cleaning and Curation**

Raw web-scraped imagery frequently includes duplicates, broken links, irrelevant photographs, low-resolution samples, collages, thumbnails, and images that do not correspond to the intended class label. The dataset was therefore cleaned through a two-stage process: an interactive cleaning stage removed obvious outliers and repeated downloads, and manual review then verified class correctness. Retained images had to show a visually recognizable example of the intended statue category with enough object information to support classification. Images were excluded when the target object was absent, the label was ambiguous, the photograph was too degraded for useful visual learning, the image was a collage or thumbnail, or the sample was a near-duplicate likely to distort the split. This manual curation step functioned as a label-governance mechanism for keeping the class taxonomy internally consistent.

For traceability, the scraping workflow can export source URLs to a CSV file so that retained samples can be audited or re-downloaded when necessary. The curated image folders were archived and synchronized across local hardware and Google Colab for version control and repeatable experimentation. This emphasis on provenance and packaging is especially important in web-sourced heritage datasets because source pages can change over time, and reproducibility depends on preserving the benchmark logic even when the raw web environment evolves.

Manual curation was especially important for classes whose names are used inconsistently across websites or whose visual appearance overlaps with broader sculptural traditions. In such cases, a purely keyword-driven crawl would inflate the dataset with semantically weak samples and make later accuracy values difficult to interpret. The review stage, therefore, served not only to remove noise but also to enforce a coherent class policy that later makes model evaluation more meaningful [Colace et al., 2025] [Paolanti et al., 2026]. The complete acquisition and preparation workflow is shown in Figure 4.1.

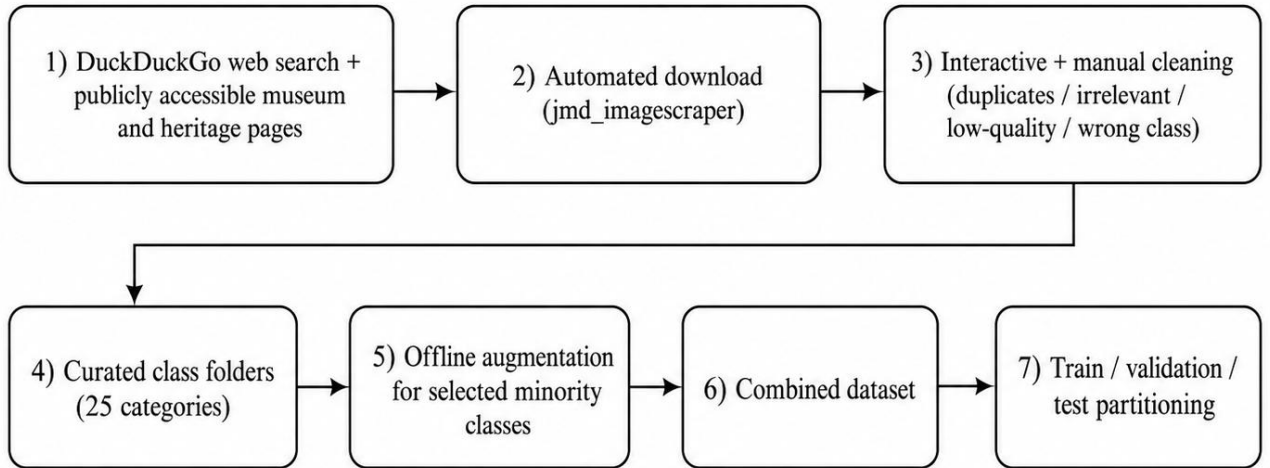


Figure 4.1: Overview of the data acquisition and preparation pipeline for EgyStatues-v1.

To ensure label reliability, the manually curated images were reviewed by a domain expert in ancient Egyptian archaeology. The expert verified that each retained image genuinely corresponded to its assigned class label and flagged ambiguous, mislabeled, or visually inconclusive samples for removal. This expert-in-the-loop validation step complements the automated and manual filtering described above and provides an additional layer of label-

quality assurance beyond the researcher's own visual inspection.

#### 4.2.2.3 Data Augmentation for Minority Classes

Despite extensive scraping, the number of usable images per class remained uneven because some statue categories are rarer online than others and because web sources differ in image quality and accessibility. To mitigate imbalance and improve robustness, additional offline augmentation was generated for selected underrepresented classes using TensorFlow/Keras ImageDataGenerator [Shorten and Khoshgoftaar, 2019] [Ottoni and Ottoni, 2025]. In the context of this thesis, it is important to distinguish this dataset-building augmentation from the model-training preprocessing used later in the benchmark chapter: the former supports dataset stabilization, whereas the latter is part of the experimental protocol and is discussed separately in Chapters 5 and 6.

The offline augmentation configuration combined several geometric transformations, including `rotation_range=90`, `width_shift_range=0.2`, `height_shift_range=0.2`, `shear_range=0.2`, `zoom_range=0.2`, and `horizontal_flip=True`, with `fill_mode` set to 'nearest'. For efficiency during generation, the temporary augmented outputs were produced at  $150 \times 150$  resolution and later resized within the classifier training pipeline to the network input size. This step broadened viewpoint and scale diversity without changing the semantic class definition of the retained images.

In the archived dataset-construction workflow, offline augmentation was performed within the class folders before the final train-validation-test partition was created. In the retained augmentation run, 20 transformed samples were generated for each of 42 source images from an underrepresented class, producing 840 augmented samples. The applied transformations included rotation, horizontal and vertical shifts, shear, zoom, and horizontal flipping. The generated images were stored at  $150 \times 150$  pixels using sequential filenames such as `augmentedImage1.jpg` and were subsequently resized to the model-specific input resolution during training. The final dataset size of 41,161 images therefore includes both curated source images and offline-augmented samples. Representative outputs are shown in Figure 4.2.

Because the final partitioning was performed after the augmented samples had been merged into the class folders, the archived workflow did not explicitly enforce source-image grouping during the split. Consequently, an original image and one or more of its transformed variants may have been assigned to different partitions. This creates a potential risk of data leakage between the training, validation, and test subsets and may result in optimistic performance estimates. Since explicit source-to-derivative lineage metadata was not retained in the archived augmented filenames, complete post-hoc verification of partition independence is not possible from the filenames alone. The reported results should therefore be interpreted with this limitation in mind.

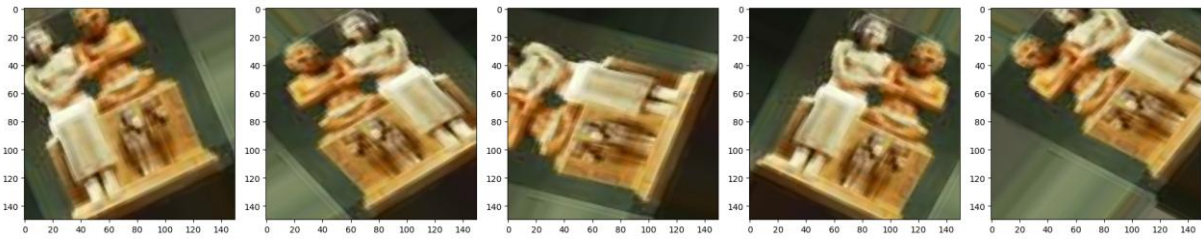


Figure 4.2: Examples of augmented samples generated using ImageDataGenerator (rotations, shifts, shear, and zoom).

#### 4.2.2.4 Dataset Packaging and Versioning

For reproducibility and ease of sharing across training environments, the curated dataset folders were archived (ZIP) and synchronized across local hardware and Google Colab. The final curated dataset (EgyStatues-v1) is maintained as a versioned research dataset, and the experimental settings are documented in the methodology and experimental chapters of the thesis. This packaging approach facilitates consistent experiments and verifiable reproduction of reported results without relying on code listings in the body of the thesis. The final list of retained statue categories is presented in Table 4.2.

Table 4.2: The 25 statue categories included in EgyStatues-v1.

| ID | Class              | ID | Class                |
|----|--------------------|----|----------------------|
| 1  | Block Statue       | 2  | Colossi of Memnon    |
| 3  | Gayer-Anderson cat | 4  | Great Sphinx of Giza |
| 5  | Head of Nefertem   | 6  | Khafre Enthroned     |

| ID | Class                            | ID | Class                                                       |
|----|----------------------------------|----|-------------------------------------------------------------|
| 7  | Khufu Statue                     | 8  | Mask of Tutankhamun                                         |
| 9  | Nefertiti Bust                   | 10 | Seneb                                                       |
| 11 | Sphinx of Memphis                | 12 | Statue of Amenemhat III                                     |
| 13 | Statue of Amun                   | 14 | Statue of Amun in the form of a ram protecting King Taharqa |
| 15 | Statue of Anubis                 | 16 | Statue of Bes                                               |
| 17 | Statue of Horus                  | 18 | Statue of Isis                                              |
| 19 | Statue of Osiris                 | 20 | Statue of Ramesses II                                       |
| 21 | Statue of Sekhmet                | 22 | Statue of the lady Tiye                                     |
| 23 | The Seated Scribe                | 24 | Thutmose                                                    |
| 25 | William the Faience Hippopotamus |    |                                                             |

#### 4.2.2.5 Dataset Split and Class Distribution

The final curated benchmark contains 41,161 images across 25 classes. This total includes the offline-augmented samples generated before the final partitioning stage. The dataset was subsequently divided into training, validation, and test subsets using the combined class folders. Therefore, the archived split should not be interpreted as a source-grouped split in which every original image and all of its transformed derivatives were guaranteed to remain within the same partition. It is partitioned into training (n=28,802; 69.97%), validation (n=6,163; 14.97%), and test (n=6,196; 15.05%) subsets. The split preserves class representation across all three partitions and provides a clear basis for controlled comparison among the retained classification models. Because the same split is reused across experiments, the benchmark supports direct interpretation of architecture differences without confounding them with data-partition changes. The detailed class-wise distribution is summarized in Table 4.3, the split-level image counts are shown in Figure 4.3, and the sorted per-class totals are

shown in Figure 4.4.

Table 4.3: Class-wise distribution across training, validation, and test splits.

| <b>Class</b>                                                | <b>Train</b> | <b>Validation</b> | <b>Test</b> | <b>Total</b> |
|-------------------------------------------------------------|--------------|-------------------|-------------|--------------|
| Block Statue                                                | 1,153        | 247               | 248         | 1,648        |
| Colossi of Memnon                                           | 1,096        | 234               | 236         | 1,566        |
| Gayer-Anderson cat                                          | 1,216        | 260               | 262         | 1,738        |
| Great Sphinx of Giza                                        | 1,272        | 272               | 274         | 1,818        |
| Head of Nefertem                                            | 1,058        | 226               | 228         | 1,512        |
| Khafre Enthroned                                            | 1,031        | 221               | 222         | 1,474        |
| Khufu Statue                                                | 1,073        | 230               | 231         | 1,534        |
| Mask of Tutankhamun                                         | 1,155        | 247               | 248         | 1,650        |
| Nefertiti Bust                                              | 1,176        | 252               | 252         | 1,680        |
| Seneb                                                       | 1,058        | 226               | 228         | 1,512        |
| Sphinx of Memphis                                           | 1,131        | 242               | 243         | 1,616        |
| Statue of Amenemhat III                                     | 1,096        | 234               | 236         | 1,566        |
| Statue of Amun                                              | 1,000        | 214               | 216         | 1,430        |
| Statue of Amun in the form of a ram protecting King Taharqa | 1,128        | 241               | 243         | 1,612        |
| Statue of Anubis                                            | 1,195        | 256               | 257         | 1,708        |
| Statue of Bes                                               | 1,220        | 261               | 263         | 1,744        |
| Statue of Horus                                             | 1,092        | 234               | 234         | 1,560        |
| Statue of Isis                                              | 1,087        | 233               | 234         | 1,554        |

| Class                            | Train  | Validation | Test  | Total  |
|----------------------------------|--------|------------|-------|--------|
| Statue of Osiris                 | 1,121  | 240        | 241   | 1,602  |
| Statue of Ramesses II            | 1,270  | 272        | 273   | 1,815  |
| Statue of Sekhmet                | 1,276  | 273        | 275   | 1,824  |
| Statue of the lady Tiye          | 1,332  | 285        | 286   | 1,903  |
| The Seated Scribe                | 1,332  | 285        | 286   | 1,903  |
| Thutmose                         | 1,121  | 240        | 241   | 1,602  |
| William the Faience Hippopotamus | 1,113  | 238        | 239   | 1,590  |
| Total                            | 28,802 | 6,163      | 6,196 | 41,161 |

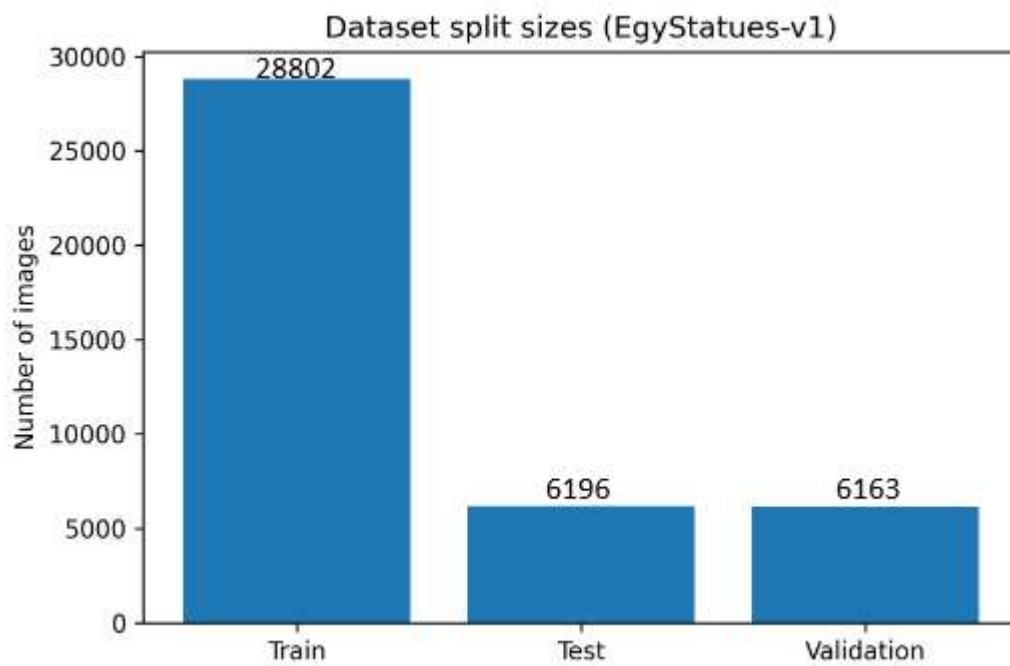


Figure 4.3: Number of images in each dataset split (training/validation/test).

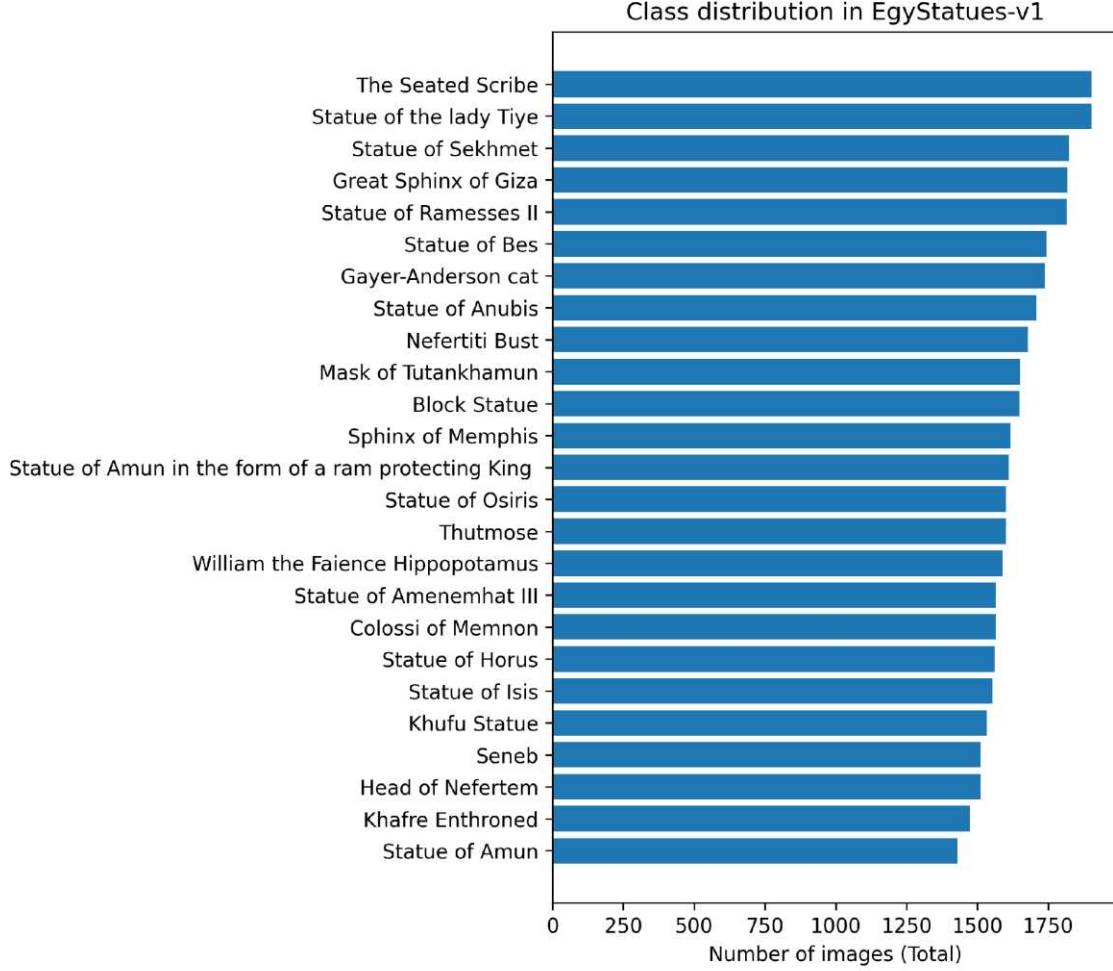


Figure 4.4: Total number of images per class in EgyStatues-v1 (sorted).

Class sizes range from 1,430 images (Statue of Amun) to 1,903 images (Statue of the lady Tiye and The Seated Scribe), indicating moderate rather than extreme imbalance. Analytically, later performance differences should not be attributed to sample count alone. A class with more images may still be difficult if it exhibits large viewpoint variation or strong overlap with neighboring categories, whereas a smaller class may remain easier when its visual identity is distinctive.

Accordingly, the dataset should be interpreted as moderately imbalanced but not severely skewed. The more important challenge is fine-grained visual adjacency: some classes are harder because they share iconographic motifs, poses, or museum-display conditions with nearby categories. Later chapters therefore interpret model behavior through a combination of support size, visual similarity, viewpoint variation, and residual source noise rather than through class counts alone [Wisniamongkol et al., 2024] [Mishra and Lourenço, 2024].

A further issue is source bias. Highly photographed statues are more likely to appear in professional museum images, whereas less famous categories may be represented through noisier web photos. Such imbalance in image quality can influence model learning even when class counts appear similar. Recognizing this possibility is important for a fair interpretation of reported accuracies [Mishra and Lourenço, 2024] [Paolanti et al., 2026].

#### **4.2.2.6 Data and Code Availability**

At the time of final thesis submission, the curated EgyStatues-v1 dataset is publicly accessible on Kaggle under the title “EgyStatues-v1” at <https://www.kaggle.com/datasets/naderemad208/statuesdatasetv2>. This statement reflects the current availability status of the dataset. Images were collected through keyword-based web queries per class label, following the workflow summarized earlier in this chapter. The thesis reports the preprocessing choices, hyperparameters, and evaluation protocol needed to understand and replicate the experimental design at the methodological level, while avoiding the inclusion of code listings in the submitted manuscript.

The dataset is intended strictly for academic research and educational use. Because images originate from third-party public web pages, copyright and usage rights remain with the original providers. Users should comply with the applicable terms of use, cite primary sources where possible, and avoid commercial redistribution of the images.

#### **4.2.2.7 Data Ethics, Provenance, and Responsible Reuse**

Since the dataset is assembled from publicly accessible web imagery, responsible reuse requires attention to provenance, copyright, and representational fairness. The thesis treats the dataset as a research benchmark rather than a redistributable image corpus for unrestricted downstream use. In a heritage context, ethical handling also requires sensitivity to how artifacts are framed, labeled, and culturally interpreted. A technically correct class prediction should therefore not be confused with full historical understanding.

If any content owner requests removal, affected samples can be taken down in future dataset versions. To support provenance auditing, the scraping workflow can export source

URLs for each downloaded image.

### **4.3 Benchmark Dataset Evaluation**

Dataset quality is assessed in terms of class coverage, balance, noise level, and representativeness of real-world conditions. Although augmentation helps reduce overfitting, remaining biases can stem from the online sources (background context, photographic style, and lighting). These factors are discussed further in Chapter 6 [Mishra and Lourenço, 2024] [Paolanti et al., 2026].

#### **4.3.1 Strengths of the Proposed Benchmark**

EgyStatues-v1 combines relatively large scale with a domain-specific focus. Rather than mixing broad artifact families, it concentrates on a coherent visual problem in which subtle differences matter. This design supports evaluation of fine-grained recognition behavior, not only coarse object discrimination. The explicit split policy and manual cleaning process also make the resulting model comparisons easier to interpret [Lamas et al., 2021] [Winterbottom et al., 2022].

#### **4.3.2 Remaining Limitations**

The benchmark remains limited by the nature of web-collected data. It does not guarantee uniform museum coverage, balanced photographic viewpoints, or equal metadata richness across classes. Some images may also be visually correlated because they originate from related online sources. These limitations do not invalidate the benchmark, but they restrict how far cross-domain generalization claims can be extended from average scores alone [Mishra and Lourenço, 2024] [Colace et al., 2025].

An additional limitation concerns the order of offline augmentation and dataset partitioning. Because augmentation preceded the final split and source-family grouping was not enforced, transformed variants of the same source image may occur across different partitions. This possibility may introduce optimistic bias into the reported evaluation and should be eliminated in future benchmark releases by splitting the curated original images first

and restricting offline augmentation exclusively to the training subset.

## 4.4 Performance Measures

Performance evaluation is a central component of any image-classification study, particularly in fine-grained cultural-heritage recognition tasks where visual similarity between classes, class imbalance, and heterogeneous image conditions can all influence the interpretation of results. For this reason, the present thesis does not rely on a single metric but instead uses a set of complementary performance measures to assess overall correctness, class-sensitive behavior, and diagnostic strengths and weaknesses across the evaluated models. The following subsection introduces the principal measures adopted throughout the experimental analysis and explains their relevance to the ancient Egyptian statue classification problem.

### 4.4.1 Common Performance Measures

Performance is evaluated using overall accuracy and class-wise metrics derived from the confusion matrix. For a given class, let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Precision measures the proportion of images predicted as belonging to a class that are correctly classified, whereas recall measures the proportion of actual images from that class that are correctly identified. The F1-score represents the harmonic mean of precision and recall and provides a balanced measure when both false-positive and false-negative errors are important [Goodfellow et al., 2016].

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$F1 - score = 2 \times (Precision \times Recall) / (Precision + Recall)$$

Overall accuracy measures the proportion of all correctly classified images among the total number of evaluated images. For a multi-class classification task, it is calculated as:

$$Accuracy = \text{Number of correctly classified images} / \text{Total number of evaluated images}$$

Although accuracy provides an intuitive summary of overall model performance, it may be influenced by class distribution because classes containing more images contribute more strongly to the final value. Therefore, additional aggregated F1 measures are used to provide a more complete evaluation.

Macro F1 is calculated by computing the F1-score independently for each class and then taking the unweighted arithmetic mean across all classes. Let  $C$  denote the number of classes and  $F1_c$  denote the F1-score of class  $c$ . Macro F1 is defined as:

$$Macro\ F1 = \frac{1}{C} \sum_{c=1}^C F1_c$$

Because every class contributes equally, Macro F1 is particularly useful for determining whether the model performs consistently across both larger and smaller classes. A low Macro F1 relative to accuracy may indicate that the model performs well on common or easier classes but poorly on some less represented or more difficult categories.

Weighted F1 also combines the class-specific F1-scores, but each class is weighted according to its support, which is the number of true samples belonging to that class. Let  $n_c$  denote the number of samples in class " $c$ ," and let " $N$ " denote the total number of evaluated samples. Weighted F1 is defined as:

$$Weighted\ F1 = \sum_{c=1}^C \frac{n_c}{N} F1_c$$

The principal difference between Macro F1 and Weighted F1 is therefore the treatment of class size. Macro F1 assigns equal importance to every class, regardless of the number of samples, whereas Weighted F1 gives greater influence to classes containing more images. Accuracy evaluates the proportion of correct predictions at the individual-image level, while Macro F1 and Weighted F1 additionally account for the balance between precision and recall. Reporting all three measures provides a more informative assessment of model performance, particularly when class frequencies and recognition difficulty differ across the twenty-five statue categories.

#### **4.4.2 Other Measures and Diagnostic Tools**

In addition to scalar metrics, confusion matrices are used to analyze which classes are frequently confused, highlighting visually similar categories and guiding future dataset refinement [Goodfellow et al., 2016].

#### **4.4.3 Detection Measures Used for YOLO Baselines**

Because the thesis also evaluates detector-style baselines, additional measures are required. In object detection, Intersection over Union (IoU) quantifies the spatial overlap between the predicted and ground-truth bounding boxes. Mean Average Precision (mAP) summarizes detection quality across classes and confidence thresholds, while precision and recall capture the trade-off between false alarms and missed detections. Reporting these measures alongside image-level accuracy makes it possible to compare model families without conflating classification and localization performance [Ali and Zhang, 2024] [Mnasri and D'Andrea, 2025].

### **4.5 Summary**

This chapter introduced the construction, curation, split organization, and evaluation measures of EgyStatues-v1. The next chapter presents the proposed framework for ancient Egyptian statue classification and describes the associated preprocessing, model-selection, and training procedures.

## Chapter 5

# Proposed Approach for Ancient Egyptian Statue Classification

### 5.1 Introduction

The proposed approach is an end-to-end statue-recognition pipeline that combines dataset curation, CNN-based transfer learning, benchmark evaluation, and cautious deployment interpretation. Its purpose is not only to maximize held-out classification accuracy, but also to establish a documented baseline that later studies can extend with multimodal inputs, transformer backbones, partial fine-tuning, or institution-specific validation.

The proposed approach follows a pragmatic design choice. Instead of introducing a highly customized architecture before stabilizing the dataset and task formulation, the thesis begins with strong pretrained CNN backbones and evaluates them under a documented benchmark setup. This choice follows directly from the preceding literature review: in heritage AI, transparent benchmarking is often more valuable than premature architectural complexity because the harder scientific problem is frequently data realism and scope control rather than raw model novelty. The resulting end-to-end workflow is shown in Figure 5.1.



Figure 5.1: Overview of the proposed end-to-end pipeline for ancient Egyptian statue classification.

Figure 5.1 summarizes the complete workflow adopted in this thesis. The first two stages—data collection, cleaning, and labeling—correspond to the construction of EgyStatues-v1 and are described in detail in Sections 4.2.2.1 and 4.2.2.2. These stages are

retained in the figure because they define the benchmark input used by the proposed classification approach. To avoid unnecessary duplication, their procedural details are not repeated in this chapter. The detailed methodological description in Chapter 5 therefore begins with input preprocessing in Section 5.2, followed by augmentation, transfer-learning model selection, and training configuration. The final evaluation stage is presented in Chapter 6.

## 5.2 Input Dataset Preprocessing

Having established the dataset-acquisition and curation stages in Chapter 4, this section describes the preprocessing applied to the retained benchmark images before model training. All images are resized to  $224 \times 224$  pixels and normalized to the  $[0, 1]$  range before being passed to the CNN classifiers. Near-duplicate images and repeated downloads were reduced during dataset construction to improve the quality of the benchmark. However, because offline augmentation preceded the final split and source-family grouping was not enforced, complete independence between the training, validation, and test partitions cannot be guaranteed.

### 5.2.1 Rationale for Input Standardization

Standardizing the input resolution to  $224 \times 224$  provides a practical compromise between visual fidelity and computational efficiency. A larger resolution might preserve finer decorative details, but it would also increase memory cost and reduce the number of controlled experiments that can be run under comparable settings. Because all compared CNN backbones can operate effectively at this scale, the choice improves comparability across models. However, the comparison remains influenced by architecture-specific heads and partly model-specific preprocessing choices.

### 5.2.2 Label Consistency and Split Integrity

Preprocessing also includes conceptual cleaning, not only pixel-level normalization. Ensuring that each image belongs to the intended category and minimizing the overlap of related or near-duplicate images are important requirements for trustworthy evaluation. In the archived benchmark, class labels were manually reviewed and repeated downloads were reduced; however, source-family grouping for offline-augmented derivatives was not enforced

during the final partitioning stage. The possibility of related variants occurring across different partitions should therefore be considered when interpreting the reported results.

### 5.2.3 Augmentation Strategy

Data augmentation was used in two related but distinct stages of the workflow. During the construction of EgyStatues-v1, offline geometric augmentation was applied to images from an underrepresented class using the TensorFlow/Keras ImageDataGenerator. The retained configuration included rotations, horizontal and vertical shifts, shear, zoom, and horizontal flipping. Twenty transformed samples were generated from each of forty-two source images, producing 840 additional images. The complete parameter configuration and representative examples are reported in Section 4.2.2.3.

The offline-augmented samples were merged into the class folders before the final train-validation-test split was created. Because source-image grouping was not enforced, transformed variants of the same source image may occur in different partitions. This potential split-integrity limitation is acknowledged in Section 5.2.2 and is considered when interpreting the experimental results. In addition, the retained model-training notebooks used partly model-specific augmentation intensities rather than one fully identical augmentation policy across all CNN backbones. The reported comparison should therefore be interpreted as a controlled transfer-learning baseline rather than as a strict augmentation-ablation study.

## 5.3 Selection of Candidate Models

The rationale for selecting the compared architectures is summarized in Table 5.1.

Table 5.1: Rationale for selecting the compared architectures in the proposed benchmarking pipeline.

| Model            | Why It Was Included                                              | Expected Strength                              |
|------------------|------------------------------------------------------------------|------------------------------------------------|
| EfficientNetV2B0 | Modern, efficient CNN with a strong transfer-learning reputation | Balanced accuracy and computational efficiency |
| ResNet50V2       | Robust residual baseline widely used in vision benchmarks        | Stable optimization and strong generalization  |

| Model             | Why It Was Included                             | Expected Strength                                                    |
|-------------------|-------------------------------------------------|----------------------------------------------------------------------|
| MobileNetV3Small  | Lightweight model for resource-aware deployment | Low computational cost                                               |
| VGG19             | Classical deep CNN baseline                     | Established conventional baseline for comparison                     |
| YOLOv11 / YOLOv12 | Detector-style reformulation of the problem     | Tests whether localization-oriented models add value in this dataset |

### 5.3.1 Architecture-Selection Rationale

The selected backbones span multiple design philosophies. EfficientNetV2B0 represents modern compound scaling with attention to efficiency. ResNet50V2 contributes a strong and well-understood residual architecture that remains a common reference point in transfer-learning studies. MobileNetV3Small represents the lightweight end of the design spectrum and is relevant for resource-constrained deployment. VGG19 serves as a classical deep CNN baseline whose relative simplicity makes overfitting behavior easier to interpret. The YOLO baselines are included not as direct competitors in the same formal ranking, but as exploratory detector-style experiments that clarify whether localization-oriented models are appropriate for this largely single-object classification problem.

To evaluate effectiveness and deployability, both high-performing and lightweight architectures were tested: EfficientNetV2B0, ResNet50V2, MobileNetV3Small, and VGG19. YOLOv11 and YOLOv12 were also tested as detection-based baselines to study cross-architecture behavior.

## 5.4 Training Configuration

The reported CNN experiments follow a shared transfer-learning framework built around ImageNet initialization, frozen feature extraction during the retained training runs, categorical cross-entropy loss, Adam-based optimization [Kingma and Ba, 2015], and validation-guided checkpoint selection. The exact monitored validation metric and classifier head were architecture-specific rather than perfectly identical across all backbones. EfficientNetV2B0 used a Dense 256-128-128 head; ResNet50V2 used global average pooling

with batch normalization and Dense 256/64 layers plus dropout; MobileNetV3Small used Dense 256/256 with dropout 0.1; and VGG19 used Dense 256/256 with batch normalization and dropout 0.1. The resulting protocol should therefore be read as a controlled feature-extraction benchmark rather than as a fully unified end-to-end fine-tuning study.

#### **5.4.1 Optimization Strategy and Model Selection**

The training strategy favors interpretability over aggressive model-specific optimization. A common split, a shared optimizer family, and documented stopping logic make the benchmark easier to read and reproduce. The thesis also makes explicit that augmentation and preprocessing were not fully standardized and remained partly model-specific in the archived notebooks. This does not invalidate the comparison, but it means that the reported benchmark should be interpreted as a strong first baseline rather than the final possible upper bound for the dataset. A stricter future comparison should rerun the leading backbones under one identical augmentation policy and with partial fine-tuning of the upper convolutional layers.

#### **5.4.2 Reproducibility and Practical Deployment**

A second design consideration is reproducibility under realistic hardware constraints. Heritage laboratories, museums, and smaller research groups may not always have access to high-end computing or large annotation teams. For this reason, the benchmark includes both relatively strong and relatively lightweight CNN families, and the thesis explicitly reports the hardware context and practical training settings. This makes the findings more useful for real deployment decisions because model choice in heritage work depends not only on peak accuracy, but also on feasibility, stability, and maintenance cost.

The final training configuration and hyperparameter summary for the evaluated models are presented in Table 5.2.

Table 5.2: Summary of training configuration and hyperparameters for all evaluated models.

| Model            | Input           | Batch | Max Epochs | Early Stop (patience) | Optimizer           | Pretrained | Notes                                                             |
|------------------|-----------------|-------|------------|-----------------------|---------------------|------------|-------------------------------------------------------------------|
| EfficientNetV2B0 | 224<br>×<br>224 | 32    | 20         | 4                     | Adam                | ImageNet   | Dense 256/128/128 head; accuracy close to the best-performing CNN |
| ResNet50V2       | 224<br>×<br>224 | 32    | 50         | 20                    | Adam                | ImageNet   | GAP + BN + Dense 256/64 + dropout 0.2; highest CNN test accuracy  |
| MobileNetV3Small | 224<br>×<br>224 | 32    | 30         | 4                     | Adam                | ImageNet   | Dense 256/256 + dropout 0.1                                       |
| VGG19            | 224<br>×<br>224 | 32    | 50         | 5                     | Adam                | ImageNet   | Dense 256/256 + BN + dropout 0.1                                  |
| YOLOv11n         | 640<br>×<br>640 | 32    | 30         | —                     | Ultralytics default | Pretrained | Detection baseline; fastest inference                             |
| YOLOv12m         | 640<br>×<br>640 | 16    | 50         | —                     | Ultralytics default | Pretrained | Detection baseline: higher mAP                                    |

### 5.4.3 Deployment Threat Model and Practical Constraints

Practical constraints also include privacy of collections, computing limits in museums or research labs, and the need for maintainable retraining cycles. Large models are not automatically better if the deployment environment cannot support them reliably. The proposed methodology is therefore deliberately pragmatic: it favors reproducible baselines, moderate complexity, and transparent comparison. This makes the thesis useful not only for

research benchmarking but also for institutions that need a clear starting point for implementation.

Any realistic deployment must anticipate failure modes. The first failure mode is distribution shift: museum images captured in-house may differ from the public web imagery used for training. The second is open-set input, where the uploaded statue belongs to a class outside the 25 benchmark categories. The third is partial visibility, strong reflections from glass cases, unusual cropping, or damaged objects that remove exactly the features most helpful for recognition. The fourth is metadata mismatch, where an image is visually plausible but paired with a misleading label. A robust heritage system should therefore be treated as a recommendation tool supported by thresholding, uncertainty handling, and expert verification.

#### **5.4.4 Model-Selection Criteria for Heritage Deployment**

The heritage domain strengthens this multi-criteria view because mistakes have interpretive consequences. A misclassification is not merely an abstract error; it can affect retrieval quality, metadata consistency, or the efficiency of downstream review. For that reason, the thesis adopts a deployment-aware interpretation of the benchmark. The best-performing model is identified, but the surrounding discussion also asks which family is best suited to constrained environments, which family is more compatible with future multimodal extension, and which family best fits single-object museum imagery. This broader framing is closer to how institutions actually choose tools.

Model selection should not be reduced to the highest validation score. In real cataloging or documentation systems, several criteria matter simultaneously: final test accuracy, training stability, computational cost, ease of maintenance, interpretability, tolerance to low-quality images, and suitability for human review. A slightly less accurate model may still be preferable if it is easier to deploy, faster to retrain, or less brittle under distribution shift. Conversely, a more accurate model may be worth the added cost if it yields a clearly better ranking of likely classes and thus reduces curator effort. These deployment-oriented criteria are summarized in Table 5.3.

### 5.4.5 Fair-Comparison Protocol and Reproducibility Safeguards

Reproducibility in heritage AI is especially important because datasets assembled from web and museum sources are difficult to reconstruct exactly. Links disappear, images are reprocessed, collection portals change, and curation judgments may drift over time. Therefore, experimental clarity must be built into the written methodology. The thesis emphasizes stable dataset packaging, explicit class definitions, preserved split logic, and transparent reporting of model settings so that a future researcher can reproduce the benchmark reasoning even if the source landscape evolves.

Fair comparison is a central design principle of the experimental protocol. In many applied deep-learning studies, apparent performance differences are confounded by unequal preprocessing, inconsistent augmentation, or hidden hyperparameter search. The present thesis reduces that risk by fixing the main benchmark elements: the same train/validation/test split, the same nominal input size for the CNN classifiers, the same loss family, and the same broad transfer-learning logic. Fairness should not be overstated, however, because the practical notebooks still used partly model-specific preprocessing and augmentation intensity. This limitation is stated explicitly so that observed ranking differences are interpreted carefully rather than as a perfectly isolated measure of architecture quality.

Table 5.3: Deployment-oriented criteria used to interpret the model families beyond raw accuracy.

| Criterion          | Why it matters in heritage use          | Preferred property  | Implication for model choice                |
|--------------------|-----------------------------------------|---------------------|---------------------------------------------|
| Accuracy           | Controls retrieval and metadata quality | High, but not alone | Used with stability and qualitative review  |
| Training stability | Reduces variance and rework             | Smooth convergence  | Favors reliable transfer-learning pipelines |
| Inference cost     | Affects deployment on modest hardware   | Moderate footprint  | Lightweight models remain relevant          |

| <b>Criterion</b>    | <b>Why it matters in heritage use</b> | <b>Preferred property</b>      | <b>Implication for model choice</b>      |
|---------------------|---------------------------------------|--------------------------------|------------------------------------------|
| Interpretability    | Supports curator trust                | Explorable errors and evidence | Encourages explanation-oriented analysis |
| Robustness to noise | Web and museum images vary strongly   | Tolerance to shift and clutter | Important for real-world adoption        |

## 5.5 Summary

This chapter described the proposed statue-classification pipeline, including input standardization, model-family selection, transfer-learning configuration, and the safeguards used to interpret the comparison cautiously. The next chapter reports the experimental setup and results, with particular attention to model behavior, class-wise performance, and the limits of deployment claims.

## Chapter 6

### Experimental Work

#### 6.1 Introduction

This chapter presents the experimental configuration and the reported benchmark results obtained using EgyStatues-v1. It evaluates the selected CNN-based transfer-learning classifiers and the exploratory YOLO detection baselines under the methodological framework described in Chapter 5. The analysis connects comparative performance, learning behavior, class-wise results, and observed errors to benchmark design, task formulation, and practical deployment constraints.

The goal of the experimental chapter is not simply to rank models. A useful result in this domain should combine recognition quality with methodological transparency, robustness to visually variable heritage imagery, and a realistic interpretation of how the system might be used in documentation workflows. Accordingly, the discussion moves from raw scores to model behavior, confusion structure, error sources, and operational implications.

#### 6.2 Hardware and Software Settings

The experimental workflow was implemented primarily in Python 3 using Jupyter-style notebooks executed on both a local workstation and Google Colab. Python was used for data preparation, image processing, model construction, training, evaluation, and visualization. The archived notebooks identify the principal software libraries used in the workflow; however, exact package-version numbers were not recorded at the time of the original experiments and therefore are not reported retrospectively.

TensorFlow and its Keras high-level API were used to construct, train, and evaluate the CNN-based transfer-learning models, including EfficientNetV2B0, ResNet50V2, MobileNetV3Small, and VGG19. Keras utilities were also used for image loading, preprocessing, callbacks, early stopping, checkpoint management, and the offline

augmentation procedure implemented through ImageDataGenerator. The CNN models were initialized using ImageNet-pretrained weights, and categorical cross-entropy was used as the classification loss.

NumPy was used for numerical array manipulation and the handling of image and prediction outputs. Matplotlib was used to generate training and validation curves, comparative performance figures, and other visualizations reported in this chapter. The Pillow library was used for image reading and basic image format handling during the dataset preparation and augmentation stages.

The data-acquisition workflow was executed in Google Colab using the open-source `jmd_imagescraper` package for DuckDuckGo-based image retrieval and interactive image cleaning. Standard Python modules, including `os`, `pathlib`, and `shutil`, were used for directory management, file organization, archiving, and movement of images between class folders.

The exploratory YOLOv11 and YOLOv12 detection experiments were implemented using the Ultralytics framework. Ultralytics was responsible for model loading, YOLO-formatted data handling, training, validation, automatic optimizer selection, and the export of detection metrics. The YOLOv12 run used the AdamW optimizer selected automatically by the framework and was executed on a Google Colab Tesla T4 GPU.

Two principal hardware environments were used. The local workstation was equipped with an 11th Generation Intel Core i5-11400F processor running at 2.60 GHz, 16 GB of RAM, and an NVIDIA GeForce RTX 3050 GPU. Google Colab provided an additional cloud-based Python environment with GPU acceleration for selected training and detection runs. The exact Colab hardware allocation could vary between sessions; therefore, a specific GPU model is reported only where it was preserved in the archived experiment output, as in the Tesla T4 used for the YOLOv12 run.

The classification experiments used RGB input images resized to  $224 \times 224$  pixels, categorical cross-entropy loss, Adam-based optimization, validation-guided early stopping, and maximum epoch limits of up to 50, depending on the model. The detailed architecture-

specific preprocessing choices, optimizer settings, batch sizes, stopping criteria, and classifier-head configurations are reported in Table 5.2 and are not repeated here.

### 6.3 Dataset Used in the Experiments

The evaluation uses the EgyStatues-v1 split of 70% training, 15% validation, and 15% testing. The dataset contains 41,161 images across 25 classes.

The test partition was not used for hyperparameter tuning and was reserved for final comparison within the archived benchmark. However, it should be treated as an internal test split rather than as an external checkpoint. Because offline augmentation was performed before dataset partitioning and source-image families were not explicitly grouped, the possibility of related transformed variants occurring across partitions cannot be excluded. Accordingly, the reported test results provide an internal comparative estimate that may be optimistic and should be interpreted alongside the stated data-leakage limitation.

### 6.4 Model Benchmarking

A controlled benchmarking logic was adopted so that differences among the compared CNN models could be interpreted with reasonable confidence. All retained classifiers were trained on the same benchmark split at the same nominal input resolution, under ImageNet-based transfer learning, and with validation-guided checkpointing. This design substantially improves comparability, even though it does not remove every source of variance in the practical notebooks.

More specifically, the CNN runs shared the same train/validation/test partition,  $224 \times 224$  inputs, categorical cross-entropy loss, and Adam-based optimization. The main differences lie in the architecture-specific heads and in partly model-specific preprocessing settings. EfficientNetV2B0 combined a Dense 256-128-128 head with the strongest geometric augmentation; ResNet50V2 used global average pooling, batch normalization, Dense 256/64 layers, dropout 0.2, and class weighting; MobileNetV3Small used Dense 256/256 with dropout 0.1 and resize-rescale preprocessing; and VGG19 used Dense 256/256 with batch normalization, dropout 0.1, and lighter augmentation. For transparency, VGG19 is retained as

a classical CNN reference, while the main interpretation emphasizes the stronger modern CNN backbones and the lightweight MobileNet option. In addition to accuracy, the evaluation pipeline exported macro metrics, classification reports, confusion matrices, and representative misclassified samples to support diagnostic interpretation. The resulting accuracy values are presented in Table 6.1 and visualized in Figure 6.1.

Table 6.1: Comparative performance of classification models on the test split (accuracy).

| Model            | Accuracy (%) |
|------------------|--------------|
| ResNet50V2       | 92.38        |
| EfficientNetV2B0 | 92.08        |
| MobileNetV3Small | 90.93        |
| VGG19            | 83.97        |

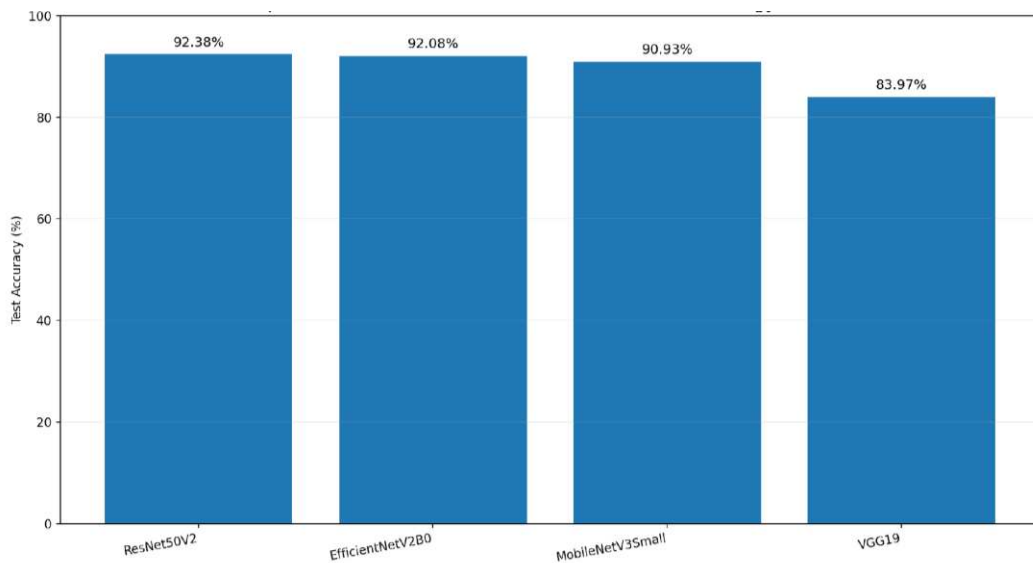


Figure 6.1: Comparative test accuracies of classification models on EgyStatues-v1.

The aggregate comparison in Table 6.1 and Figure 6.1 provides the overall ranking of the retained CNN classifiers. A more detailed analysis is presented in Section 6.5, where Figures 6.2–6.16 are discussed alongside the learning behavior and qualitative predictions of the individual CNN models. Figures 6.17–6.20 present the exploratory YOLO detection results, while Figures 6.21 and 6.22 provide representative error examples and the confusion matrix of the best-performing classifier. The YOLO experiments are interpreted separately because their task formulation and evaluation protocol differ from those of the retained CNN classification benchmark.

## 6.5 Results and Discussion

The experimental results are discussed using both quantitative and qualitative evidence. Final test metrics provide an overall comparison among the evaluated models, while the training and validation curves help assess convergence behavior, training stability, generalization gaps, and possible indications of overfitting. These dynamics are particularly relevant to fine-grained heritage recognition, where visually discriminative features, such as crown shape, animal association, posture, and local carving details, may occupy only a small region of the image.

Qualitative prediction examples complement the numerical evaluation by illustrating model behavior on individual statue images, including visually ambiguous, low-quality, and externally sourced examples. These figures help determine whether the observed errors are semantically understandable and whether model performance remains plausible across different image conditions. The following subsections therefore discuss each model together with its associated learning curves, test-set predictions, and qualitative inference examples.

### 6.5.1 EfficientNetV2B0 Results

EfficientNetV2B0 achieved a test accuracy of 92.08%, with macro-averaged precision, recall, and F1-score of 92.15%, 91.95%, and 91.96%, respectively. These results place it very close to ResNet50V2, which achieved the highest test accuracy in the retained CNN comparison. The small difference between the two models indicates that EfficientNetV2B0 should be regarded as a comparably strong baseline rather than as a substantially weaker alternative. Its learning behavior is presented through the training and validation loss and accuracy curves in Figures 6.2 and 6.3. Representative test-set predictions and a qualitative web-image inference example are shown in Figures 6.4 and 6.5, respectively.

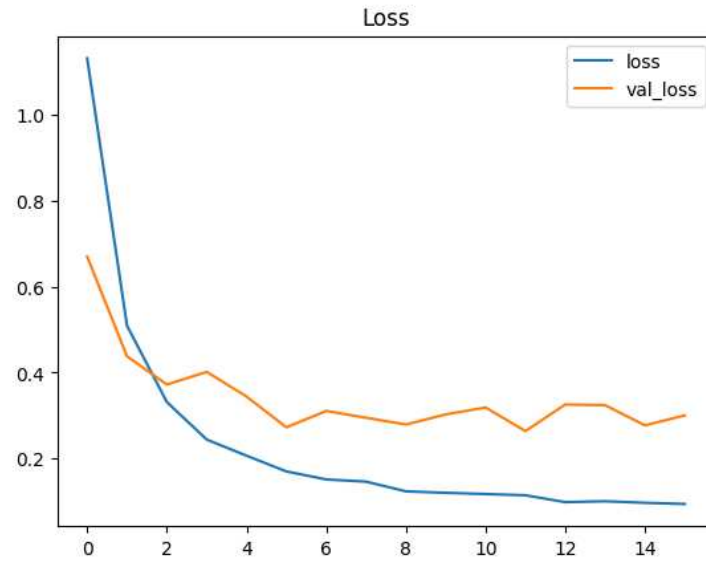


Figure 6.2: EfficientNetV2B0 training and validation loss across epochs.

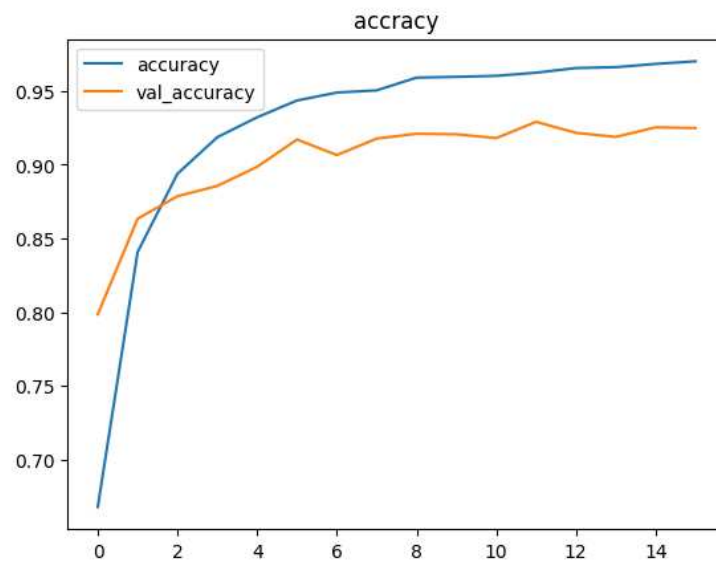


Figure 6.3: EfficientNetV2B0 training and validation accuracy across epochs.

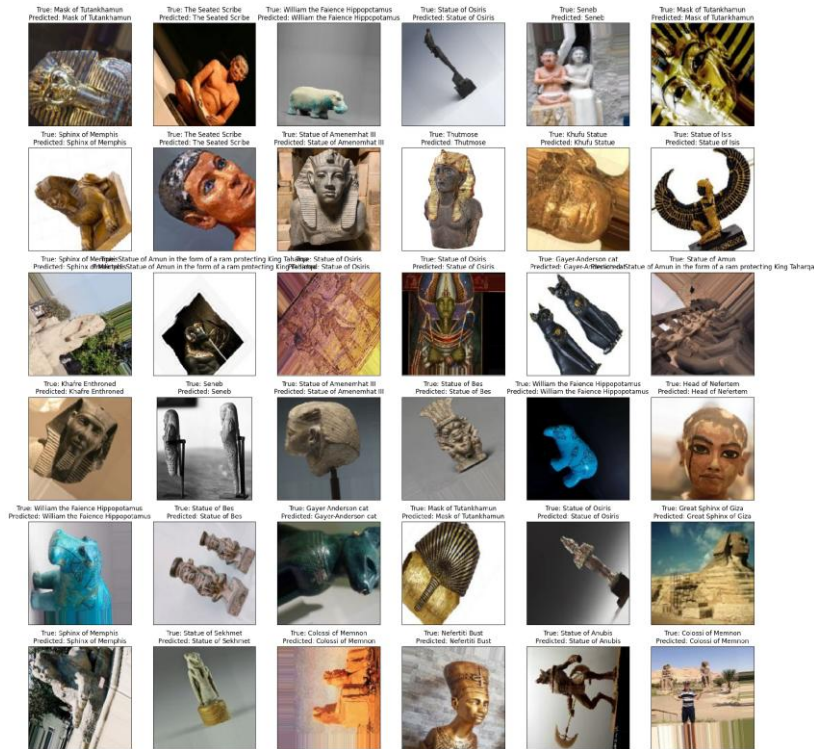


Figure 6.4: Qualitative test-set predictions of EfficientNetV2B0 (random samples with true vs predicted labels).



Figure 6.5: Single-image inference example using EfficientNetV2B0 (web image predicted as Statue of Anubis).

### 6.5.2 ResNet50V2 Results

ResNet50V2 was trained under the same split logic and input size used for the other CNN models. The ImageNet-pretrained backbone was frozen, followed by global average

pooling, batch normalization, and a compact Dense 256/64 head with dropout 0.2. On the held-out test set, ResNet50V2 achieved 92.38% accuracy, with macro-averaged precision, recall, and F1-score of 92.33%, 92.27%, and 92.28%, respectively. This is the highest final test result in the CNN comparison; however, its advantage over EfficientNetV2B0 is only marginal, so both models should be interpreted as comparably strong baselines rather than as evidence of a large architectural separation. Its loss and accuracy curves are shown in Figures 6.6 and 6.7, while qualitative predictions and web-image inference are illustrated in Figures 6.8 and 6.9.

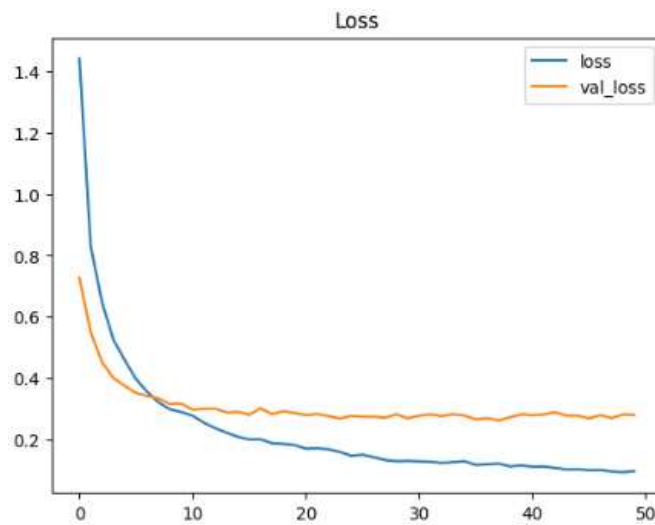


Figure 6.6: ResNet50V2 training and validation loss across epochs.

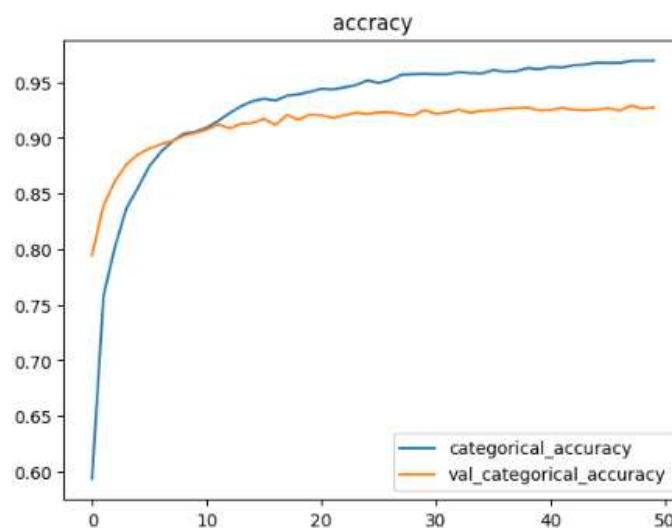


Figure 6.7: ResNet50V2 training and validation accuracy across epochs.



Figure 6.8: Qualitative test-set predictions of ResNet50V2 (random samples with true vs predicted labels).

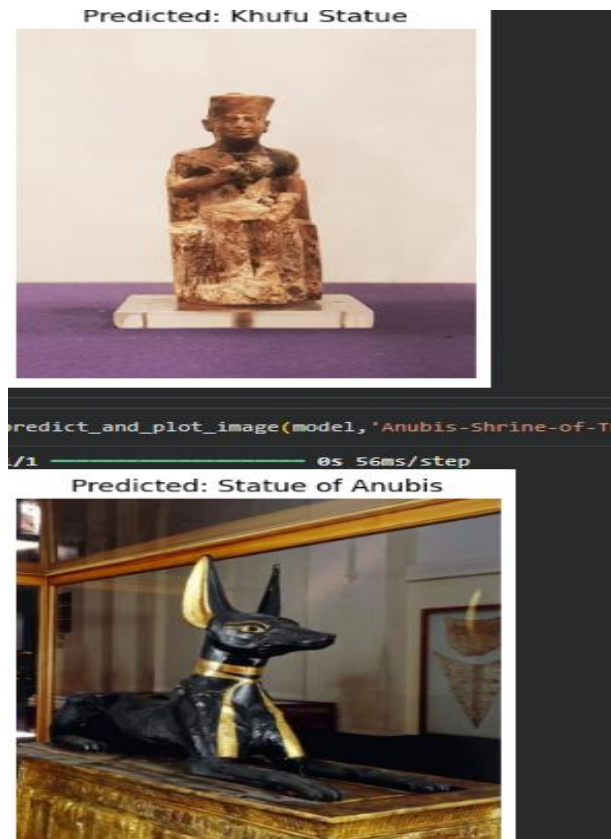


Figure 6.9: Single-image inference examples using ResNet50V2 (web images predicted as Khufu Statue and Statue of Anubis).

### 6.5.3 MobileNetV3Small Results

MobileNetV3Small was evaluated under the same split protocol and  $224 \times 224$  input resolution used for the other classifiers. Its frozen ImageNet backbone was paired with a lightweight Dense 256/256 head and dropout 0.1, which makes the model relevant for resource-constrained deployment. On the held-out test set, MobileNetV3Small achieved 90.93% accuracy, with macro precision, recall, and F1-score of 91.08%, 90.89%, and 90.87%, respectively. This places it below ResNet50V2 and EfficientNetV2B0, but still within a

practically strong range for museum-oriented or educational support workflows, provided that future external validation confirms similar behavior. Its learning curves are shown in Figures 6.10 and 6.11, while qualitative predictions and single-image inference are shown in Figures 6.12 and 6.13.

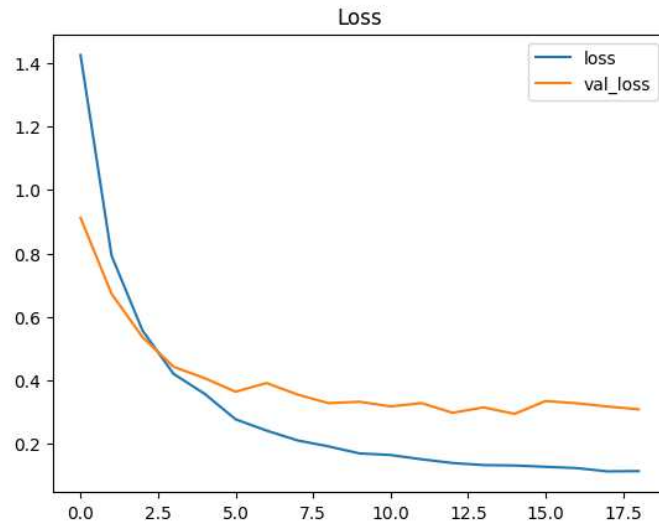


Figure 6.10: MobileNetV3Small training and validation loss across epochs.

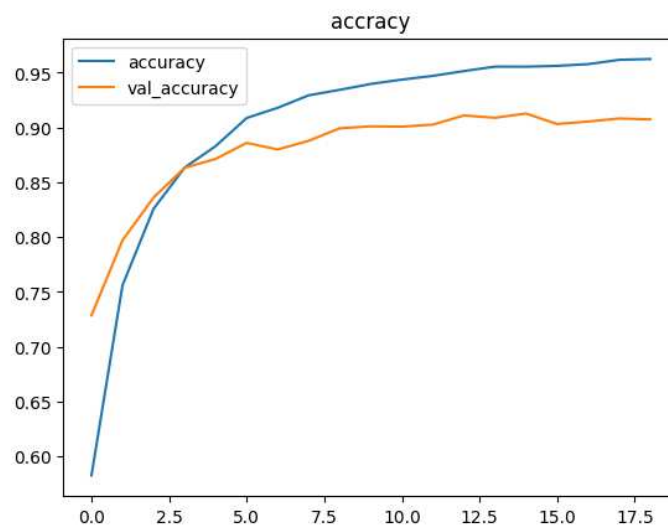


Figure 6.11: MobileNetV3Small training and validation accuracy across epochs.

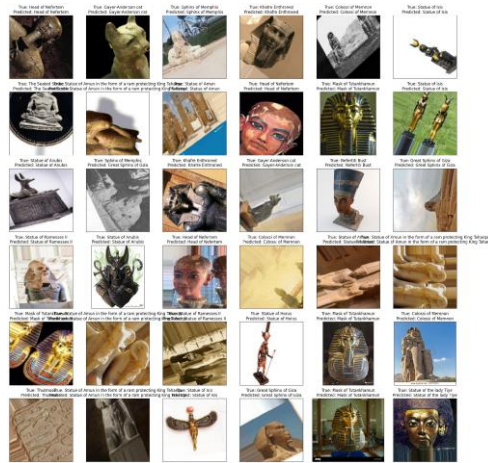


Figure 6.12: Qualitative test-set predictions of MobileNetV3Small (random samples with true vs predicted labels).



Figure 6.13: Single-image inference example using MobileNetV3Small (web image predicted as Statue of Anubis).

### 6.5.4 VGG19 Results

VGG19 was trained with the same input size and broad transfer-learning logic, but it remained weaker than the more modern CNN backbones. On the held-out test set, VGG19 achieved 83.97% accuracy, with macro precision, recall, and F1-score of 84.03%, 83.86%, and 83.72%, respectively. This result confirms that VGG19 remains a useful classical CNN reference, but it still lags behind ResNet50V2, EfficientNetV2B0, and MobileNetV3Small in this fine-grained statue-recognition benchmark. The VGG19 learning behavior is shown in Figure 6.14, and its qualitative predictions are illustrated in Figures 6.15 and 6.16.

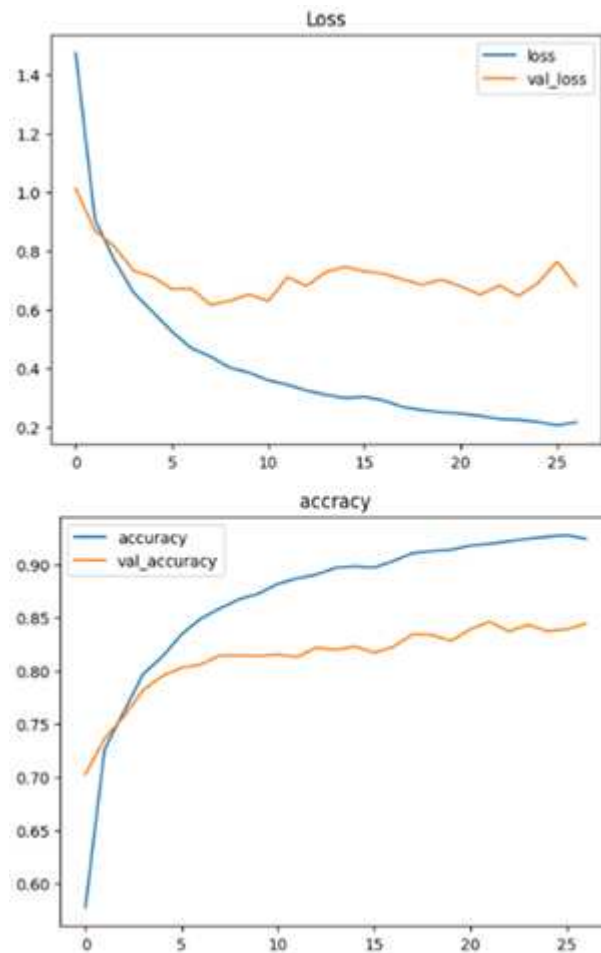


Figure 6.14: VGG19 training and validation loss/accuracy across epochs.

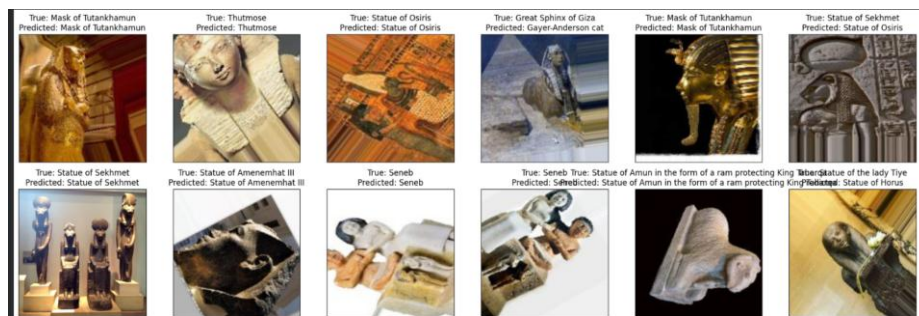


Figure 6.15: Qualitative test-set predictions of VGG19 (random samples with true vs predicted labels).

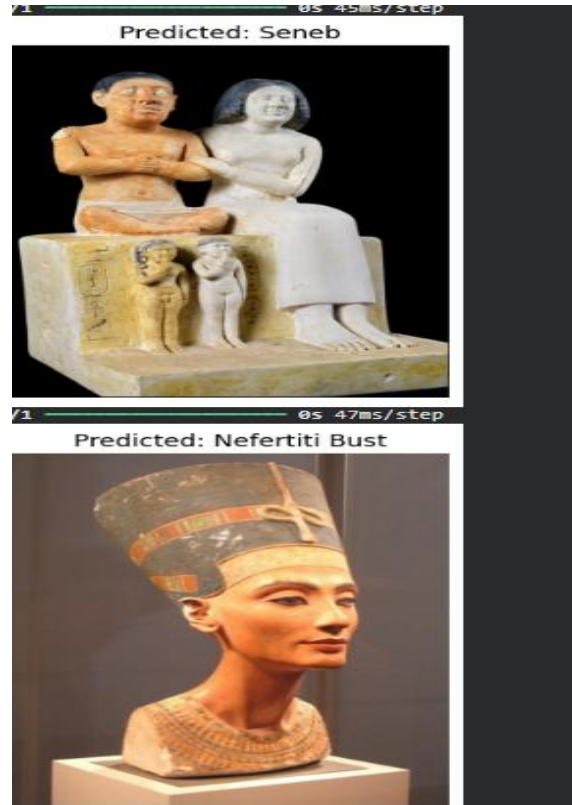


Figure 6.16: Single-image inference examples using VGG19 (web images predicted as Seneb and Nefertiti Bust).

Across the final CNN comparison, ResNet50V2 delivered the highest held-out test performance (92.38%), followed very closely by EfficientNetV2B0 (92.08%), MobileNetV3Small (90.93%), and VGG19 (83.97%). The same ordering is broadly supported by the macro precision, recall, and F1-score values, which suggests that the ranking is not simply an artifact of easier high-frequency classes. The narrow gap between ResNet50V2 and EfficientNetV2B0 should be treated as a marginal difference, while MobileNetV3Small remains the strongest lightweight deployment-oriented option among the tested classifiers.

### 6.5.5 Why Image Classification Was More Suitable Than Detection in This Setting

Although the CNN classifiers achieved strong performance within the retained classification benchmark, their accuracy values should not be compared directly with the mAP values of the YOLO detectors because the two experimental settings used different task formulations, data formats, subsets, and evaluation metrics. Nevertheless, the experimental evidence indicates that image classification was the more suitable primary formulation for the current dataset. Most images contain one dominant and approximately centered statue, making

explicit object localization less necessary. Under these conditions, a classification model can focus directly on fine-grained category cues, whereas a detector must jointly address localization and category prediction even when localization provides limited additional value.

This finding should not be interpreted as evidence that detection models are generally inferior for cultural-heritage applications. Detection remains appropriate when an image contains multiple artifacts, when the object occupies only a small region of the scene, or when the location of architectural elements, inscriptions, or surface defects is itself required. In the present study, the YOLOv11 and YOLOv12 experiments are therefore treated as exploratory detection baselines that clarify the effect of task formulation rather than as directly comparable entries in the formal CNN classification ranking. Their detailed results are presented in Sections 6.5.6 and 6.5.7.

### 6.5.6 YOLOv12 Results (Detection Baseline)

To examine whether detection-oriented architectures could still offer useful comparative insight, a YOLOv12m model was fine-tuned using the Ultralytics framework on a YOLO-formatted subset of the dataset. This experiment should be read as an exploratory side study rather than as an alternative formal benchmark entry, because the underlying data format, objective, and evaluation logic differ from those of the retained classification comparison. Even so, the detector run remains informative because it illustrates what is gained and lost when the same heritage problem is reframed around localization.

The YOLO training used a YOLO-formatted dataset with 4,895 training images and 2,263 validation images (one annotated instance per image). Ultralytics' automatic optimizer selection chose AdamW ( $\text{lr}=0.000345$ ), and training completed in approximately 4.09 hours on a Google Colab Tesla T4 GPU. Overall detection performance on the validation split reached precision = 0.793, recall = 0.599,  $\text{mAP}@0.5$  = 0.688, and  $\text{mAP}@0.5:0.95$  = 0.688, with an average inference time of 15.3 ms per image (plus 1.2 ms post-processing). The overall comparison is presented in Table 6.2 and shown in Figure 6.17, while the lowest-performing YOLOv12 classes are summarized in Table 6.3 and Figure 6.18.

Table 6.2: Overall detection performance of YOLO baselines on the validation split.

| Model    | Precision (%) | Recall (%) | mAP@0.5 (%) | mAP@0.5:0.95 (%) | Inference (ms) |
|----------|---------------|------------|-------------|------------------|----------------|
| YOLOv12m | 79.3%         | 59.9       | 68.8        | 68.8             | 15.3           |
| YOLOv11n | 73.5%         | 59.7       | 65.5        | 65.4             | 4.1            |

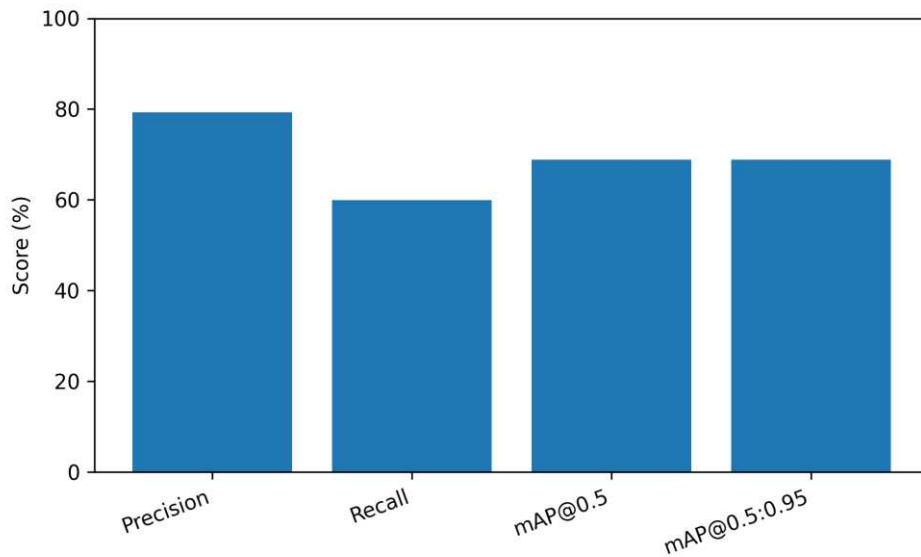


Figure 6.17: YOLOv12 overall detection metrics on the validation split.

Table 6.3: Five lowest-performing classes under YOLOv12 by mAP@0.5 (validation split).

| Class                                                       | Images | Precision (%) | Recall (%) | mAP@0.5 (%) |
|-------------------------------------------------------------|--------|---------------|------------|-------------|
| Statue of Amun in the form of a ram protecting King Taharqa | 63     | 28.0          | 11.1       | 19.5        |
| Sphinx of Memphis                                           | 49     | 84.0          | 20.4       | 33.3        |
| Statue of Amun                                              | 62     | 82.6          | 30.6       | 44.2        |
| Seneb                                                       | 41     | 100.0         | 28.6       | 46.3        |
| Khufu Statue                                                | 29     | 82.9          | 33.5       | 47.3        |

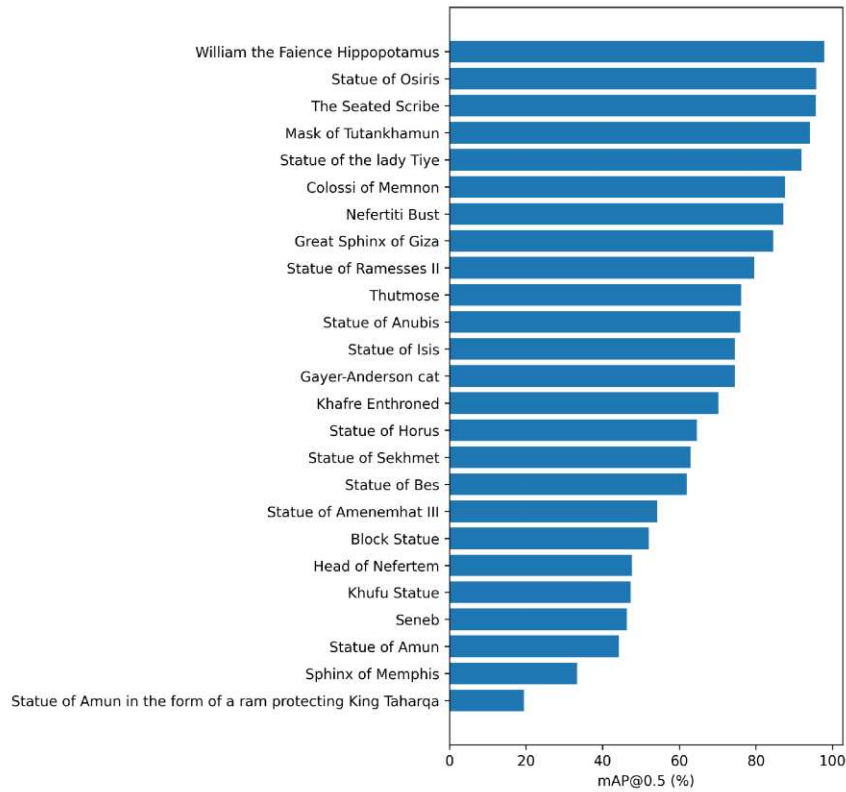


Figure 6.18: Per-class mAP@0.5 for YOLOv12 on the validation split.

Although YOLOv12 achieved reasonable detector-level performance, the detector formulation remained less aligned with the thesis setting, where each image typically contains a single centered artifact and the target is category-level recognition rather than bounding-box localization. Its value in the thesis is therefore interpretive: it helps demonstrate why a detector-style formulation is not the most natural primary benchmark for this dataset.

### 6.5.7 YOLOv11 Results (Detection Baseline)

A YOLOv11n detector was also trained as a lightweight exploratory baseline using the same YOLO-formatted subset. Like the YOLOv12 experiment, this run is retained for methodological completeness and deployment-oriented discussion, not because it belongs in the formal ranking of the classification study. Within the detection formulation, YOLOv11 provides faster inference, whereas YOLOv12 achieves higher detection metrics. Together with the predominantly single-object structure of the dataset, these findings reinforce the decision to retain image classification as the primary task formulation. The weakest YOLOv11 classes are summarized in Table 6.4, and the YOLOv11/YOLOv12 comparison and YOLOv11 class-level profile are shown in Figures 6.19 and 6.20.

Table 6.4: The five lowest-performing YOLOv11 classes on the validation split, ranked from the lowest to the highest mAP@0.5.

| Class                                                                | Images | Precision (%) | Recall (%) | mAP@0.5 (%) |
|----------------------------------------------------------------------|--------|---------------|------------|-------------|
| Statue of Amun<br>in the form of a<br>ram protecting<br>King Taharqa | 63     | 21.4          | 7.9        | 13.1        |
| Sphinx of<br>Memphis                                                 | 49     | 64.2          | 22.0       | 29.6        |
| Khufu Statue                                                         | 29     | 81.1          | 29.7       | 37.1        |
| Seneb                                                                | 41     | 69.6          | 27.9       | 37.4        |
| Statue of Amun                                                       | 62     | 55.9          | 32.7       | 39.0        |

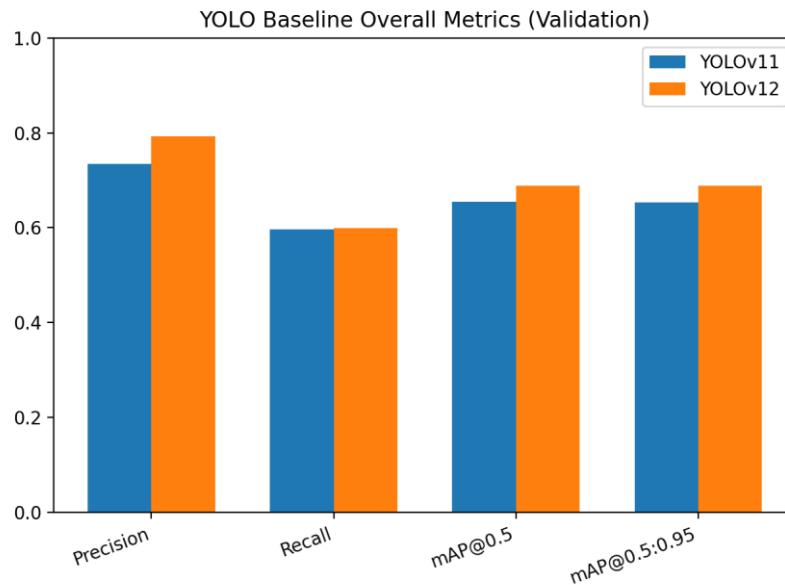


Figure 6.19: Comparison of YOLOv11 vs YOLOv12 overall detection metrics on the validation split.

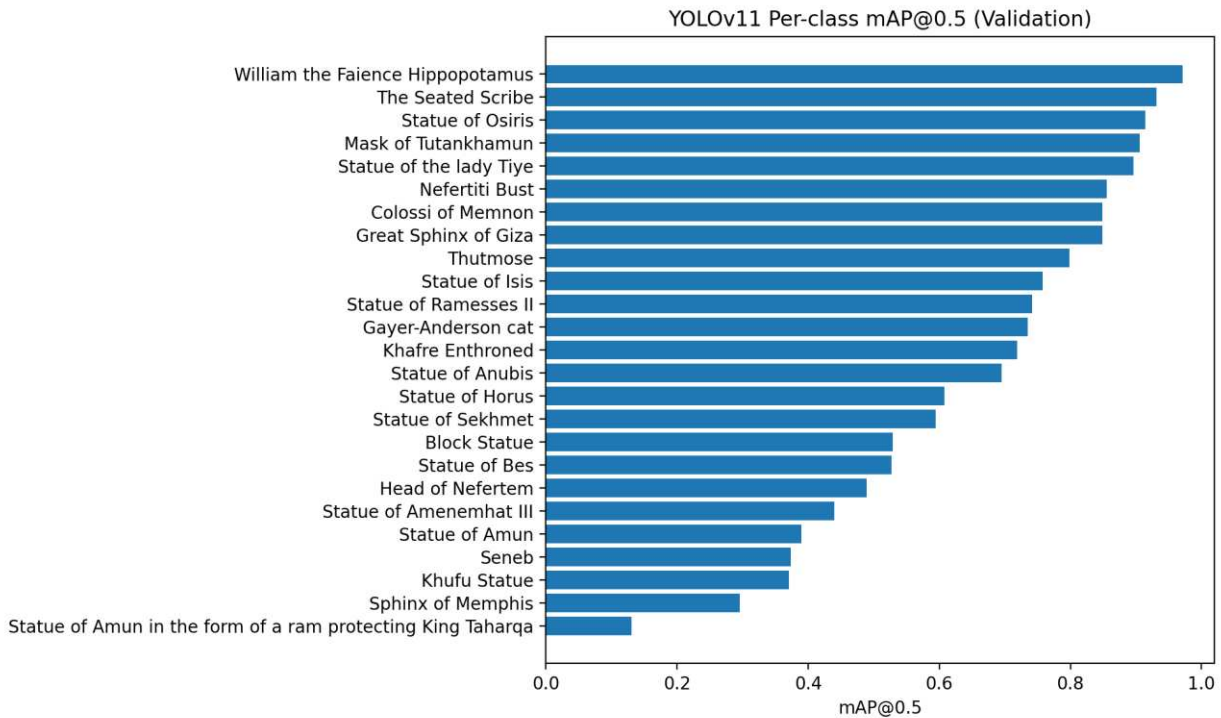


Figure 6.20: Per-class mAP@0.5 for YOLOv11 on the validation split.

### 6.5.8 Per-class Analysis and Relation to Data Distribution

Although the final dataset is only moderately imbalanced across the 25 classes, class-wise performance still varies meaningfully. The aggregate metrics confirm this point. ResNet50V2 achieved 92.33% macro precision, 92.27% macro recall, and 92.28% macro F1-score, slightly ahead of EfficientNetV2B0 (92.15%, 91.95%, and 91.96%) and MobileNetV3Small (91.08%, 90.89%, and 90.87%). VGG19 reached 84.03% macro precision, 83.86% macro recall, and 83.72% macro F1-score, confirming that it remains useful as a classical CNN reference but weaker than the newer architectures. Because the gap between accuracy and macro F1 remains small for the three strongest CNNs, their performance is not being carried by only a handful of easy classes. These macro-averaged results are summarized in Table 6.5.

The confusion matrix of the best final classifier, ResNet50V2, shows that the hardest mistakes are semantically plausible rather than arbitrary. The strongest recurring confusion occurs between Statue of Amun and Statue of Amun in the form of a ram protecting King Taharqa, with 43 Statue of Amun test images predicted as the Taharqa ram-protection class. Smaller recurrent confusions also appear among visually adjacent categories such as Thutmose

and Statue of Amenemhat III, Khafre Enthroned and Statue of Ramesses II, and Horus-related versus Anubis-related classes. These errors reflect shared silhouettes, similar frontal compositions, overlapping royal or divine iconography, and cases where decisive details such as headdress shape, animal association, or local carving cues occupy only a small region of the image. These confusion patterns are shown in Figure 6.22.

At the same time, some categories were recognized very reliably. Under EfficientNetV2B0, classes such as Head of Nefertem, Seneb, Statue of the Lady Tiye, Mask of Tutankhamun, and William the Faience Hippopotamus showed recall values in the mid- to high-90% range. The weaker classes were typically those with close visual neighbors or context-sensitive presentation, especially the Statue of Amun, Khafre Enthroned, and the Taharqa ram-protection class. This pattern suggests that visual proximity, viewpoint variation, and local cue visibility are more influential than mild class imbalance alone.

Table 6.5: Macro-averaged precision, recall, and F1-score of the classification models on the held-out test split.

| <b>Model</b>     | <b>Test Accuracy (%)</b> | <b>Macro P (%)</b> | <b>Macro R (%)</b> | <b>Macro F1 (%)</b> |
|------------------|--------------------------|--------------------|--------------------|---------------------|
| ResNet50V2       | 92.38                    | 92.33              | 92.27              | 92.28               |
| EfficientNetV2B0 | 92.08                    | 92.15              | 91.95              | 91.96               |
| MobileNetV3Small | 90.93                    | 91.08              | 90.89              | 90.87               |
| VGG19            | 83.97                    | 84.03              | 83.86              | 83.72               |

The representative misclassified samples in Figure 6.21 reinforce the same interpretation. Many mistakes arise under oblique viewpoints, partial cropping, glare, damaged surfaces, or a museum-display context that hides the decisive iconographic cue. Importantly, most wrong predictions remain within a visually neighboring class rather than jumping to a completely unrelated category, which suggests that the learned representation captures broad statue families even when fine-grained distinctions remain difficult.

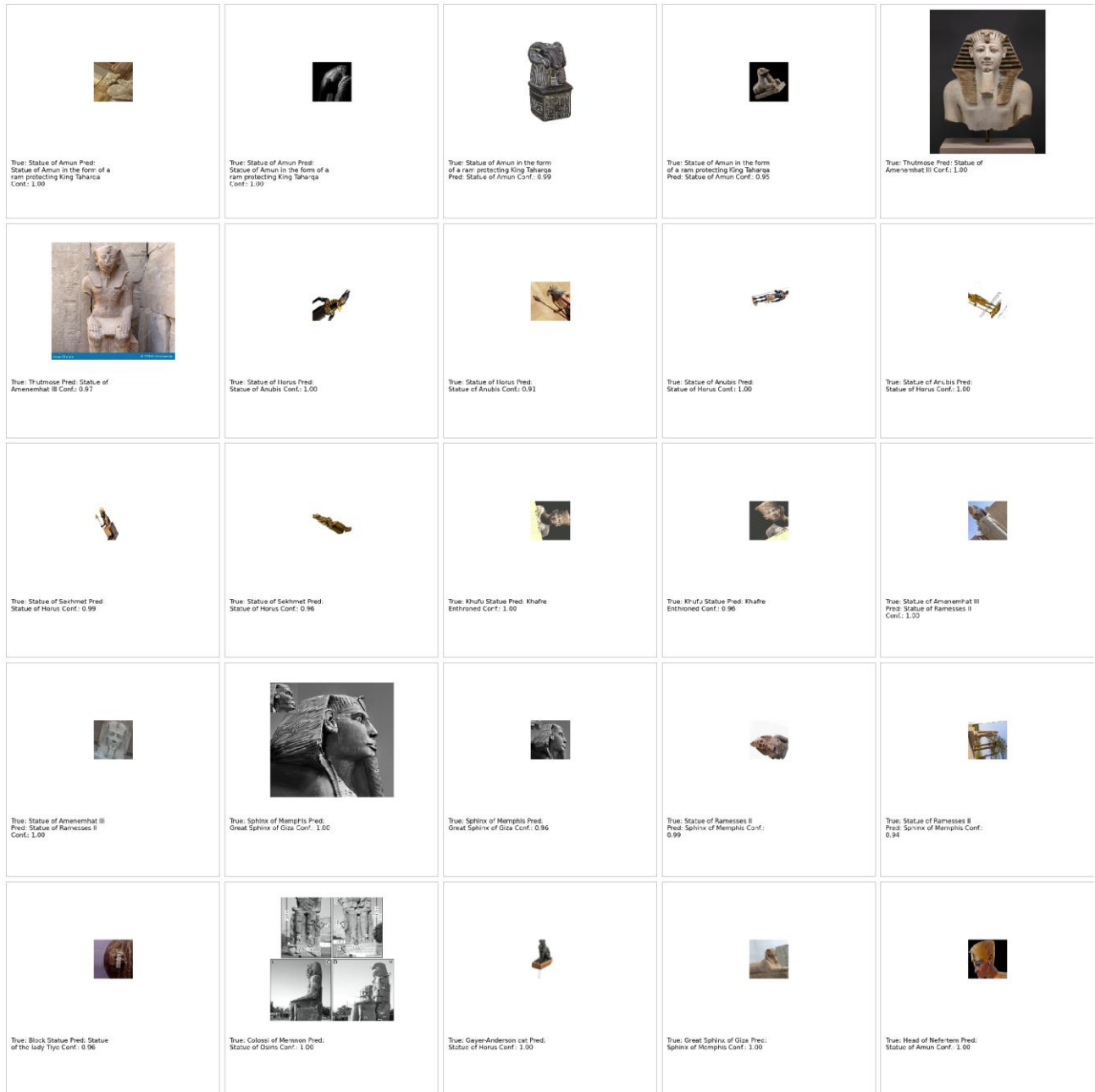


Figure 6.21: Representative misclassified test samples for ResNet50V2.

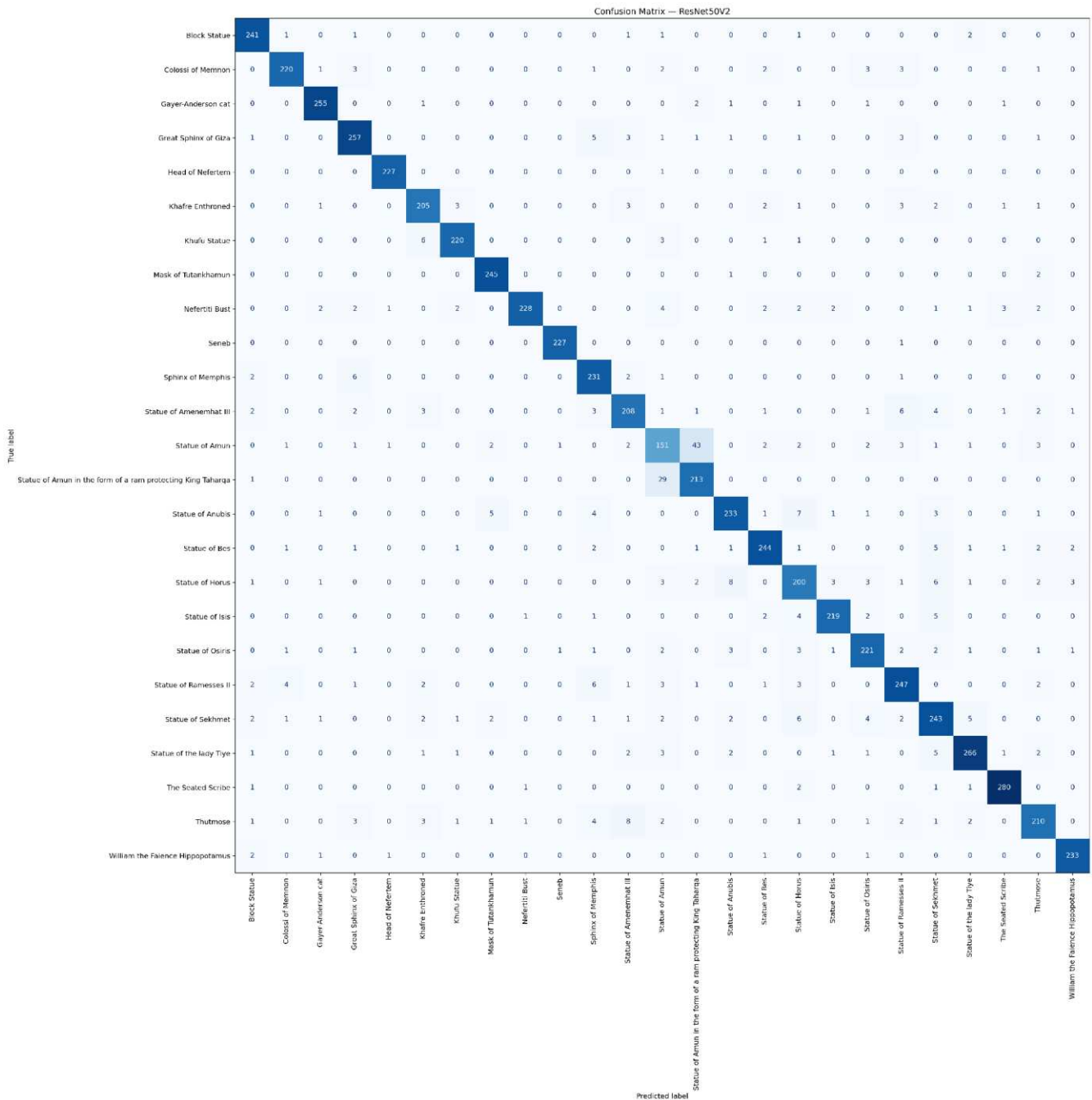


Figure 6.22: Confusion matrix of ResNet50V2 on the held-out test split.

### 6.5.9 Relation to the Broader Literature

The experimental findings are consistent with the broader literature synthesized in the review component of the thesis. Recent cultural-heritage AI studies repeatedly show that benchmark quality, transfer learning, and evaluation realism can matter as much as architectural novelty. The present thesis demonstrates this pattern in the specific context of ancient Egyptian statues: with a curated taxonomy, fixed splits, and transparent CNN-based

transfer learning, strong recognition results can be obtained before moving to more complex multimodal or geometry-aware extensions.

### **6.5.10 Practical Implications for Cataloging and Museum Use**

A realistic deployment scenario would treat the classifier as an assistive layer rather than an autonomous authority. For example, the model could propose the top few likely categories for a newly uploaded image, support keyword enrichment in digital archives, help route uncertain cases for curator review, or power educational retrieval interfaces that connect visually similar objects. In such settings, even imperfect predictions can be valuable if they are ranked sensibly, accompanied by calibrated confidence, and embedded within expert-supervised workflows.

### **6.5.11 Threats to Validity and Remaining Limitations**

Several limitations remain. First, the dataset is assembled from public web sources, so subtle source correlations and collection-style biases may remain even after cleaning and split control. Second, the benchmark is still image-only and does not yet exploit metadata, inscriptions, provenance information, or 3D shape cues that could help resolve difficult cases. Third, the retained formal benchmark is based on internal train/validation/test partitioning plus a qualitative external probe, not yet on a large independently curated institutional testbed. Fourth, because the reported CNNs use frozen backbones and partly model-specific preprocessing, the present ranking should be read as a conservative baseline rather than as an optimized upper bound or a definitive architecture ranking.

An additional threat to validity concerns the order of offline augmentation and dataset partitioning. In the archived workflow, augmented samples were merged into the class folders before the final train-validation-test split was created, and source-family grouping was not enforced. Consequently, an original image and one or more visually related transformed variants may occur in different partitions. This potential leakage may produce optimistic evaluation estimates and limits the strength of conclusions that can be drawn from the internal test results. A future benchmark release should first partition the curated original images and then restrict offline augmentation exclusively to the training subset.

A further threat to validity is background correlation. If a class is photographed disproportionately in one museum, against one pedestal style, or beside a recurring display card, the model may learn contextual shortcuts that are only indirectly related to the statue itself. For this reason, strong performance should be interpreted as evidence of useful visual discrimination within the benchmark setting, not as a guarantee of universal robustness across all future collections.

These limitations define the strength of the thesis claims. The work offers a literature-grounded problem framing, a reusable benchmark dataset, and a controlled CNN-based baseline for fine-grained statue recognition. It does not claim universal generalization, multimodal understanding, or production-ready institutional deployment. The contribution should therefore be read as a documented benchmark foundation with clearly stated methodological boundaries.

Accordingly, the most defensible conclusion is that the proposed pipeline provides a useful but limitation-aware internal benchmark and remains promising as an assistive research tool, provided that its numerical results are interpreted cautiously and that future releases address source-family separation before any stronger deployment claims are made.

### **6.5.12 Recommended Human-in-the-Loop Workflow**

A practical deployment workflow should keep the model in a human-in-the-loop role. Rather than assigning an uncontestable label, the system should return a ranked shortlist of likely classes together with confidence values and, where possible, a compact explanation of the visual evidence. A curator, archivist, or trained researcher can then verify the prediction, correct it when necessary, and log difficult cases for future dataset refinement. This preserves scholarly authority while reducing routine search and cataloging effort.

The same workflow fits the practical lessons of the experiments. Classification-focused pipelines are well matched to single-object, centered imagery common in digital documentation, whereas detector-style baselines may become more useful in future scenarios involving multiple artifacts, cluttered scenes, or mixed display environments.

### 6.5.13 Robustness, External Validity, and Generalization

Robustness should be interpreted relative to the data distribution and the limitations of the archived partitioning workflow. The internal test split provides a comparative estimate of model behavior on the retained dataset; however, because offline augmentation preceded partitioning and source-family grouping was not enforced, it cannot provide strong evidence of leakage-free internal validity. The best classifier also produced encouraging qualitative behavior on additional web images and photographs of real statues that were not included in the train, validation, or test partitions. Nevertheless, this qualitative probe is not a substitute for a fully designed and independently curated external benchmark.

For that reason, the present results support a cautious claim of promising practical generalization, not a definitive claim of universal robustness. A broader controlled external benchmark remains an important next step, especially if the intended use extends from academic experimentation toward cataloging support, digital search, or public-facing heritage access.

### 6.5.14 Error Analysis and Plausible Sources of Confusion

Error analysis provides a more informative view of model behavior than aggregate performance measures alone. The confusion matrix, class-wise results, and representative misclassified samples indicate that several errors arise from visually similar categories, incomplete visual evidence, and variations in image acquisition conditions. In addition to these observed patterns, practical deployment may introduce further risks, including out-of-class inputs and low-quality mobile captures that are not adequately represented in the current benchmark.

Table 6.6 summarizes the principal observed and anticipated error patterns, their plausible causes, their potential operational consequences, and the corresponding mitigation strategies. The table should be interpreted as a deployment-oriented synthesis of the experimental findings rather than as evidence that every listed mitigation strategy has already been empirically validated.

Table 6.6: Operational interpretation of common error types and recommended mitigation actions.

| Error pattern              | Likely cause                        | Operational risk                  | Recommended mitigation                           |
|----------------------------|-------------------------------------|-----------------------------------|--------------------------------------------------|
| Visually similar classes   | Local cues too small or ambiguous   | Wrong catalog suggestion          | Increase close-up diversity and use top-k review |
| Background shortcut        | Spurious museum/display correlation | Poor transfer to new institutions | Audit predictions and diversify sources          |
| Partial or damaged object  | Missing discriminative evidence     | Low-confidence mismatch           | Allow abstention and manual escalation           |
| Out-of-class statue        | Open-set input                      | Forced wrong label                | Use confidence thresholds and unknown handling   |
| Low-quality mobile capture | Blur, glare, crop shift             | Reduced retrieval value           | Add robustification data and input checks        |

The mitigation actions in Table 6.6 translate the identified failure modes into practical directions for benchmark refinement and system deployment. Increasing source and viewpoint diversity may reduce reliance on background correlations, while confidence thresholds, abstention mechanisms, and human review can limit the operational impact of uncertain or out-of-class predictions. These measures remain recommendations for future evaluation and should be validated through controlled experiments before being treated as confirmed improvements.

## 6.6 Summary

Overall, the experiments show that CNN-based transfer learning is the strongest formulation tested for the current single-object statue-classification task, with ResNet50V2 and EfficientNetV2B0 performing very closely and MobileNetV3Small offering the best lightweight option. The YOLO baselines remain useful as exploratory evidence about task formulation, but they are not the retained primary benchmark. The deployment-oriented model recommendations are summarized in Table 6.7.

Table 6.7: Suggested model families for different practical deployment scenarios derived from the experimental findings.

| Deployment Scenario                  | Most Suitable Model Family     | Reason                                                                           |
|--------------------------------------|--------------------------------|----------------------------------------------------------------------------------|
| Highest-accuracy offline cataloging  | ResNet50V2 / EfficientNetV2B0  | Best overall recognition quality in the benchmark; the gap between them is small |
| Resource-constrained application     | MobileNetV3Small               | Competitive accuracy with lower computational burden                             |
| Complex scenes with multiple objects | YOLO-style detector            | Localization becomes meaningful when scenes are not single-object                |
| Human-in-the-loop archive enrichment | CNN classifier + expert review | Fast shortlist generation with manageable review workload                        |

The experiments indicate the practical value of modern CNN transfer-learning backbones for fine-grained statue classification and reinforce the importance of benchmark quality, label discipline, and transparent evaluation.

Beyond ranking model families by accuracy, the experiments show that dataset curation, split integrity, and error interpretation are as important as architecture choice. The strongest CNNs benefited from the fact that the task is predominantly a single-object recognition problem in which the statue occupies the semantic center of the image. Under these conditions, transfer learning provided a reliable mechanism for adapting mature visual representations to a culturally specialized domain without requiring excessively complex training infrastructure.

The YOLO runs still played a useful methodological role because they clarified the boundary between classification and localization. Their weaker results should not be read as a general failure of detection-based systems; rather, they show that detection becomes more meaningful when scenes contain multiple artifacts, cluttered compositions, or spatially distributed points of interest. For museum cataloging images that usually foreground one statue, direct classification remains the more economical and empirically justified primary design choice.

Taken together, the findings support a practical recommendation for future heritage-AI studies: begin with a transparent benchmark, document class definitions rigorously, analyze confusion groups carefully, and only then extend the pipeline toward partial fine-tuning, transformer baselines, multimodal fusion, or institution-facing interfaces. This progression preserves scientific clarity while creating a realistic path from benchmark research to curated deployment.

## **Chapter 7**

### **Conclusion and Future Work**

This chapter consolidates the scientific contributions of the thesis, highlights their practical value for cultural-heritage documentation, and relates the work to its three connected research components: literature review, dataset construction, and benchmark evaluation.

#### **7.1 Thesis Contributions**

This thesis makes four principal contributions. First, it provides a focused synthesis of recent deep-learning research in cultural heritage and translates recurring lessons concerning data quality, transfer learning, explainability, task formulation, and deployment realism into design requirements for ancient Egyptian statue recognition.

Second, it introduces EgyStatues-v1, a curated benchmark containing 41,161 images across 25 ancient Egyptian statue categories. The dataset contribution includes a documented class taxonomy, acquisition and cleaning workflow, fixed training-validation-test organization, class-wise distribution analysis, and an explicit account of provenance, augmentation, and split-integrity limitations.

Third, it establishes a controlled CNN-centered transfer-learning benchmark using EfficientNetV2B0, ResNet50V2, MobileNetV3Small, and VGG19. Within the retained internal evaluation, ResNet50V2 achieved the highest test accuracy, EfficientNetV2B0 performed nearly identically, and MobileNetV3Small provided the strongest lightweight alternative. The comparison is interpreted cautiously because the backbones were frozen, preprocessing was partly model-specific, and offline augmentation preceded the final split.

Fourth, the thesis provides a deployment-oriented interpretation of the results. The exploratory YOLOv11 and YOLOv12 experiments clarify why direct image classification is better aligned with a predominantly single-object, category-level benchmark, while the

confusion analysis, human-in-the-loop workflow, and threat model identify safeguards required before museum-facing use.

From a software-engineering perspective, the overall contribution is therefore not a single trained model, but a structured and reusable benchmark workflow that connects versioned data preparation, documented model configuration, diagnostic evaluation, and limitation-aware deployment reasoning. The thesis does not claim universal generalization or production readiness; instead, it provides a transparent foundation for future leakage-free benchmark releases, partial fine-tuning, transformer and multimodal extensions, and controlled institutional validation.

## 7.2 Future Work

Future work should proceed in prioritized stages. The immediate next step is stronger external validation on museum-provided imagery or previously unseen web domains, together with quantitative analysis of the hardest confusion pairs. The next methodological stage should test partial fine-tuning of the upper CNN layers, ablation studies on augmentation and input resolution, and a stricter rerun under one unified augmentation policy. The benchmark should also be reconstructed by splitting the curated original images before augmentation and restricting all offline-augmented derivatives to the training subset. These steps directly address the main limits of the current benchmark: frozen backbones, partly model-specific preprocessing, and qualitative rather than fully quantitative external validation.

Future work should also continue integrating explainability, calibration, uncertainty handling, and bias analysis into the evaluation pipeline. Heritage collections are culturally significant, and any automated system should remain auditable, interpretable, and suitable for expert oversight rather than being treated as a black-box substitute for curatorial judgment.

A further direction is expansion beyond the current architectural scope. Transformer-based classifiers such as ViT, DeiT, and Swin, as well as multimodal museum-artifact models, should be evaluated after the CNN baseline has been rerun under stricter standardization. Such extensions would test whether the current CNN-centered ranking persists under broader model

families or whether the task benefits from more global visual representations and additional metadata.

### **7.3 Strategic Roadmap for Future Benchmark Releases and Deployment**

A practical roadmap can be organized into three phases. Phase one should stabilize the benchmark through versioned dataset releases, stricter duplicate control, clearer annotation guidance, and reproducible reporting templates. Phase two should extend the benchmark scientifically through partial fine-tuning studies, transformer and multimodal baselines, uncertainty-aware evaluation, and controlled external validation on institution-specific image collections. Phase three should address deployment through curator-facing interfaces, confidence thresholds, human correction loops, and periodic re-evaluation as new collections are added.

This phased structure keeps future work tied to dataset transparency and methodological comparability rather than allowing expansion to become a set of disconnected experiments. It also provides a realistic path for turning the thesis benchmark into a reusable research asset rather than a one-time experiment.

## References

### **[Ali and Zhang, 2024]**

M. L. Ali and Z. Zhang. The YOLO Framework: A comprehensive review of evolution, applications, and benchmarks in object detection. *Computers*, 13(12), 336. <https://doi.org/10.3390/computers13120336>, 2024.

### **[Barucci et al., 2021]**

A. Barucci, C. Cucci, M. Franci, M. Loschiavo and F. Argenti. A Deep Learning Approach to Ancient Egyptian Hieroglyphs Classification. *IEEE Access*, 9, 123438-123447. <https://doi.org/10.1109/ACCESS.2021.3110082>, 2021.

### **[Belhi et al., 2021]**

A. Belhi, H. O. Ahmed, T. Alfaqheri, A. Bouras, A. H. Sadka and S. Foufou. Study and evaluation of pre-trained CNN networks for cultural heritage image classification. In *Data Analytics for Cultural Heritage: Current Trends and Concepts*, pp. 47-69. Springer. [https://doi.org/10.1007/978-3-030-66777-1\\_3](https://doi.org/10.1007/978-3-030-66777-1_3), 2021.

### **[Boutet and Hallot, 2025]**

S. Boutet and P. Hallot. Ground sampling distance as a key parameter for automatic crack detection in built heritage: A practical framework with YOLOv5. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-M-9-2025, 179-186. <https://doi.org/10.5194/isprs-archives-XLVIII-M-9-2025-179-2025>, 2025.

### **[Casillo et al., 2025]**

M. Casillo, F. Colace, R. Gaeta, A. Lorusso and M. Pellegrino. Artificial intelligence in archaeological site conservation: Trends, challenges, and future directions. *Journal of Computer Applications in Archaeology*, 8(1), 224-241. <https://doi.org/10.5334/jcaa.207>, 2025.

**[Chen et al., 2025]**

X. Chen, J. He and S. Wang. Deep learning-driven pathology detection and analysis in historic masonry buildings of Suzhou. *npj Heritage Science*, 13, 197. <https://doi.org/10.1038/s40494-025-01783-y>, 2025.

**[Colace et al., 2025]**

F. Colace, R. Gaeta, A. Lorusso, M. Pellegrino and D. Santaniello. New AI challenges for cultural heritage protection: A general overview. *Journal of Cultural Heritage*, 75, 168-193. <https://doi.org/10.1016/j.culher.2025.07.019>, 2025.

**[Dosovitskiy et al., 2021]**

A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

**[Dragomir and Dragomir, 2025]**

O. E. Dragomir and F. Dragomir. AI-based proactive maintenance for cultural heritage conservation: A hybrid neuro-fuzzy approach. *Future Internet*, 17(11), 510. <https://doi.org/10.3390/fi17110510>, 2025.

**[Du and Wang, 2024]**

L. Du and Y. Wang. Bi-YOLO: A novel object detection network and dataset for components of China heritage buildings. *Journal of Building Engineering*, 97, 110817. <https://doi.org/10.1016/j.jobe.2024.110817>, 2024.

**[Duque Domingo et al., 2020]**

J. Duque Domingo, J. Gómez-García-Bermejo and E. Zalama. Egyptian Shabtis Identification by Means of Deep Neural Networks and Semantic Integration with Europeana. *Applied Sciences*, 10(18), 6408. <https://doi.org/10.3390/app10186408>, 2020.

**[Fiorucci et al., 2020]**

M. Fiorucci, M. Khoroshiltseva, M. Pontil, A. Traviglia, A. Del Bue and S. James. Machine learning for cultural heritage: A survey. *Pattern Recognition Letters*, 133, 102-108, 2020.

**[Gao et al., 2024]**

C. Gao, Q. Zhang, Z. Tan, G. Zhao, S. Gao, E. Kim and T. Shen. Applying optimized YOLOv8 for heritage conservation: Enhanced object detection in Jiangnan traditional private gardens. *Heritage Science*, 12, 31. <https://doi.org/10.1186/s40494-024-01144-1>, 2024.

**[Gokak et al., 2025]**

S. Gokak, P. Khichade, N. Heggalagi, A. Billowria and S. Hegde. Defect detection and classification of cultural heritage buildings using deep learning. In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT)*, Volume 3, pp. 619-626. SciTePress. <https://doi.org/10.5220/0013633700004664>, 2025.

**[Goodfellow et al., 2016]**

I. Goodfellow, Y. Bengio and A. Courville. *Deep Learning*. MIT Press, 2016.

**[Hasan et al., 2025]**

M. Hasan, M. K. M. Rabby, I. Jahan, M. J. A. Soeb and M. F. Jubayer. The evolution and advancement of YOLO algorithms in object detection: From real-time breakthroughs to modern architectures. *Preprints.org*. <https://doi.org/10.20944/preprints202510.2019.v1>, 2025.

**[Hassan et al., 2023]**

M. A. Hassan, A. Hamdy and M. Nasr. Egypt Monuments Dataset version 1: A Scalable Benchmark for Image Classification and Monument Recognition. *International Journal of Advanced Computer Science and Applications*, 14(4). <https://doi.org/10.14569/IJACSA.2023.0140497>, 2023.

**[He et al., 2016]**

K. He, X. Zhang, S. Ren and J. Sun. Deep residual learning for image recognition. In *Proceedings of CVPR*, 2016.

**[Howard et al., 2019]**

A. Howard et al. Searching for MobileNetV3. In *Proceedings of ICCV*, pp. 1314-1324, 2019.

**[Kingma and Ba, 2015]**

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. ICLR, 2015.

**[Lamas et al., 2021]**

A. Lamas et al. MonuMAI: Dataset, deep learning pipeline and citizen science based app for monumental heritage taxonomy and classification. Neurocomputing, 420, 266-280. <https://doi.org/10.1016/j.neucom.2020.09.041>, 2021.

**[Lu et al., 2026]**

Y. Lu, J. Li, L. Li and C. Yuan. A museum artifact classification model based on cross-modal attention fusion and generative data augmentation. Scientific Reports, 16, 298. <https://doi.org/10.1038/s41598-025-29671-2>, 2026.

**[Mishra and Lourenço, 2024]**

M. Mishra and P. B. Lourenço. Artificial intelligence-assisted visual inspection for cultural heritage: State-of-the-art review. Journal of Cultural Heritage, 65, 536-550. <https://doi.org/10.1016/j.culher.2024.01.005>, 2024.

**[Mnasri and D'Andrea, 2025]**

Z. Mnasri and A. D'Andrea. Automatic inventory of archaeological artifacts based on object detection and classification using deep and transfer learning. Digital Applications in Archaeology and Cultural Heritage, 39, e00458. <https://doi.org/10.1016/j.daach.2025.e00458>, 2025.

**[Murat and Kiran, 2025]**

A. A. Murat and M. S. Kiran. A comprehensive review on YOLO versions for object detection. Engineering Science and Technology, an International Journal, 70, 102161. <https://doi.org/10.1016/j.jestch.2025.102161>, 2025.

**[Ottoni and Ottoni, 2025]**

A. L. C. Ottoni and L. T. C. Ottoni. A deep learning approach for cultural heritage building classification using transfer learning and data augmentation. Journal of Cultural Heritage, 74, 214-224. <https://doi.org/10.1016/j.culher.2025.06.010>, 2025.

**[Paolanti et al., 2026]**

M. Paolanti, E. Frontoni and R. Pierdicca. Towards trustworthy AI in cultural heritage. *npj Heritage Science*, 14, 131. <https://doi.org/10.1038/s40494-026-02403-z>, 2026.

**[Radford et al., 2021]**

A. Radford et al. Learning transferable visual models from natural language supervision. *ICML*, 2021.

**[Rei et al., 2023]**

L. Rei et al. Multimodal metadata assignment for cultural heritage artifacts. *Multimedia Systems*, 29, 847-869. <https://doi.org/10.1007/s00530-022-01025-2>, 2023.

**[Selvaraju et al., 2017]**

R. R. Selvaraju et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of ICCV*, 2017.

**[Shorten and Khoshgoftaar, 2019]**

C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 60, 2019.

**[Simonyan and Zisserman, 2015]**

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

**[Siountri and Anagnostopoulos, 2023]**

K. Siountri and C. Anagnostopoulos. The classification of cultural heritage buildings in Athens using deep learning techniques. *Heritage*, 6(4), 3673-3705. <https://doi.org/10.3390/heritage6040195>, 2023.

**[Tan and Le, 2021]**

M. Tan and Q. V. Le. EfficientNetV2: Smaller models and faster training. *ICML*, 2021.

**[Vijayakumar and Vairavasundaram, 2024]**

A. Vijayakumar and S. Vairavasundaram. YOLO-based object detection models: A review and its applications. *Multimedia Tools and Applications*, 83(35), 83535-83574. <https://doi.org/10.1007/s11042-024-18872-y>, 2024.

**[Wen et al., 2024]**

J. Wen, X. Ma and X. Wang. Enhancing museum experience through deep learning and multimedia technology. *Heliyon*, 10(14), e34076. <https://doi.org/10.1016/j.heliyon.2024.e34076>, 2024.

**[Weyand et al., 2020]**

T. Weyand, A. Araujo, B. Cao and J. Sim. Google Landmarks Dataset v2: A large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2572-2581. <https://doi.org/10.1109/CVPR42600.2020.00265>, 2020.

**[Winterbottom et al., 2022]**

T. Winterbottom, A. Leone and N. Al Moubayed. A deep learning approach to fight illicit trafficking of antiquities using artefact instance classification. *Scientific Reports*, 12, 13468. <https://doi.org/10.1038/s41598-022-15965-2>, 2022.

**[Wisnammongkol et al., 2024]**

J. Wisnammongkol, T. Sanpechuda, S. Lim and L. Kovavisaruch. Imbalanced classification of cultural heritage images using deep neural networks. In *2024 First International Conference on Pioneering Developments in Computer Science & Digital Technologies (IC2SDT)* (pp. 46-51). IEEE, 2024.

**[Xu et al., 2025]**

W. Xu et al. ViT-HVE: A vision transformer-based framework for recognition and weighted evaluation of cultural heritage values. *npj Heritage Science*, 13, 571. <https://doi.org/10.1038/s40494-025-02143-6>, 2025.

## الملخص العربي

تتناول هذه الرسالة استخدام الذكاء الاصطناعي للتعرف على التماثيل المصرية القديمة من خلال ثلاثة مكونات مترابطة: مراجعة أساليب التعلم العميق في مجال التراث الثقافي، وإنشاء مجموعة بيانات مرجعية متخصصة (v1-EgyStatues)، وإجراء دراسة تجريبية مضبوطة لتصنيف التماثيل. وتستهدف الرسالة التصنيف الدقيق متعدد الفئات للتماثيل المصرية القديمة انطلاقاً من صور ثنائية الأبعاد بنظام RGB، بما يدعم التوثيق الرقمي وفهرسة المتاحف والاسترجاع الموجّه لعلم المصريات. ولإتاحة تقييم موثق، جُمعت بيانات v1-EgyStatues من صفحات المتاحف المتاحة للجمهور ومستودعات التراث وعمليات استرجاع الصور من الويب الموجهة بالفئات، ثم نُقيت من خلال ضبط التكرارات والتحقق اليدوي. وتضم النسخة المرجعية المعتمدة 41,161 صورة موزعة على 25 فئة، ومقسمة إلى مجموعات للتدريب (28,802) والتحقق (6,163) والاختبار (6,196). وفي جميع تجارب المقارنة الخاصة بشبكات CNN، وُحِدَت الصور إلى حجم  $224 \times 224$  بكسل، ودُرِّبَت النماذج وفق بروتوكول موثق لنقل التعلم باستخدام شبكات أساسية مُدرّبة مسبقاً على ImageNet مع تجميد أوزانها، وإيقاف مبكر موجّه بأداء مجموعة التحقق. وفي التقييم النهائي على مجموعة الاختبار المحجوزة، حقق ResNet50V2 أعلى دقة (92.38%)، تلاه EfficientNetV2B0 (92.08%)، ثم MobileNetV3Small (90.93%)، وأخيراً VGG19 (83.97%). وإلى جانب الدقة الإجمالية، تعرض الرسالة مقاييس محسوبة بالمتوسط الكلي، وتحليلاً على مستوى الفئات، وتفسيراً لمصفوفة الارتباك، وعينات تمثيلية للتصنيفات الخاطئة لتحديد مواطن الخطأ المتبقية. كما تُناقش تجارب YOLOv11 و YOLOv12 الاستكشافية بوصفها تجارب جانبية مرتبطة بصياغة المهمة، بينما يركز المعيار الرسمي المعتمد على مصنفات نقل التعلم القائمة على شبكات CNN. وتشير دراسة نوعية تكميلية باستخدام صور واقعية لتماثيل وصور إضافية من خارج التقسيم إلى إمكانات عملية واعدة، إلا أن الاستنتاج يظل متحفظاً عمداً لأن المعيار يعتمد على شبكات أساسية مجمدة، وإعدادات معالجة مسبقة خاصة جزئياً بكل نموذج، وتحقق خارجي نوعي بدلاً من تقييم كمي كامل. وتتمثل المساهمة الرئيسية في معيار موثق وإطار للتقييم يمكن البناء عليه في أبحاث الذكاء الاصطناعي متعدد الوسائط للتراث، والضبط الدقيق الجزئي، وتوسعات المحولات، والنشر الموثوق في بيانات المتاحف.

الجامعة المصرية للتعلم الإلكتروني - الأهلية  
كلية الحاسبات وتكنولوجيا المعلومات  
برنامج ماجستير هندسة البرمجيات

## منهجيات تصنيف التماثيل المصرية القديمة باستخدام التعلم العميق

رسالة مقدمة استكمالاً لمتطلبات الحصول على درجة ماجستير هندسة البرمجيات

مقدمة من الطالب  
نادر عماد يعقوب حنا  
مُدَرِّس مساعد، كلية الحاسبات وتكنولوجيا المعلومات،  
الجامعة المصرية للتعلم الإلكتروني

تحت إشراف

د. ردينة عبد السلام  
مدرس، كلية الحاسبات وتكنولوجيا  
المعلومات، الجامعة المصرية للتعلم  
الإلكتروني

د. ياسر عبد الحميد  
أستاذ مساعد، كلية الدراسات العليا  
للبحوث الإحصائية، قسم علوم الحاسب،  
جامعة القاهرة

القاهرة - مصر

2026