



Investigation Into Deep Learning Methods for a Formal Analysis of Ancient Sculptures

RESEARCH ARTICLE

ubiquity press

BIRGIT BERGMANN

DANIEL MATUSZCZYK

DENIS FISSELER

THORALF SCHRÖDER

NELE SCHRÖDER-GRIEBEL

FRANK WEICHERT

*Author affiliations can be found in the back matter of this article

ABSTRACT

The formal analysis of ancient sculptures plays a crucial role in understanding historical contexts, cultures, and societies. In-depth archaeological studies often rely on mostly manual, time-consuming and comprehensive visual analysis conducted by specialists in the field. This work explores the integration of deep learning methods, based on neural network architectures, for segmenting facial components and classifying portraits of the first four Roman emperors of the Julio-Claudian dynasty, with the aim of supporting archaeologists in their work. The segmentation is evaluated objectively using the Intersection over Union metric (IoU), with scores around 0.85, in combination with qualitative visual analysis via image masks. For the classification task, F1-scores of approximately 0.78 are achieved for emperor recognition, and 0.79 for portrait type classification. Furthermore, highlighting the decisive regions used by the network for classification offers a new perspective on the importance of facial regions compared to the traditionally considered hair patterns.

CORRESPONDING AUTHOR:

Frank Weichert

Department of Computer
Science VII, Technical
University Dortmund, 44227
Dortmund, Germany
frank.weichert@tu-dortmund.de

KEYWORDS:

ancient sculptures; typology;
style; formal analysis; emperor
portraits; deep learning

TO CITE THIS ARTICLE:

Bergmann, B, Matuszczyk, D, Fisseler, D, Schröder, T, Schröder-Griebel, N and Weichert, F. 2026. Investigation Into Deep Learning Methods for a Formal Analysis of Ancient Sculptures. *Journal of Computer Applications in Archaeology*, 9(1): 400–417. DOI: <https://doi.org/10.5334/jcaa.238>

1 INTRODUCTION

Analyzing the form of ancient sculptures plays a crucial role in understanding the art, culture and society of Greek and Roman antiquity, as it reflects a range of contextual factors, including the time and place of creation, the available technological knowledge, and the intentions or influence of the creator(s) as well as commissioner(s), to name but a few relevant factors. Because of this, formal analysis yields valuable insights not only into the artifact itself but also, on a larger scale, into the culture that produced and used it (Borbein 2000; Kunze 2015; von den Hoff 2019: 22–23). This approach is particularly indispensable in cases where contextual information is scarce or entirely lacking – conditions that frequently apply to ancient sculptures. Therefore, analyzing and grouping these artifacts based on the similarity or dissimilarity of their external form is frequently the principal (if not the sole) method of ascertaining the date and also the place of production; and in the case of

portraiture, it may even enable the identification of the individual depicted (Borbein 2000: 114, 124).

For this reason, every study of ancient sculptures begins with an analysis of their external form, on a macro- and a micro-scale: typology and style (Borbein 2000: 110–111; von den Hoff 2019: 23–27) (see Figure 1). Typology is concerned with groups or sets of forms – or one could say patterns – that were used to meet a given representational need or solve a given representational problem. Such types are e.g. the kouros for the depiction of a young man in Archaic times, the combination of a seated woman with her servant standing in front of her on Classical Attic grave reliefs or a certain arrangement of locks in the portraits of Roman emperors (Borbein 2000: 110, 120–122; von den Hoff 2019: 25–26). Style, by contrast, pertains to the formal qualities of the work itself – specifically, to the manner in which individual elements such as a lock of hair, a fold of fabric, or a muscle are rendered (for example, in a plastic or flat, rounded or angular, highly detailed or more schematic fashion), as

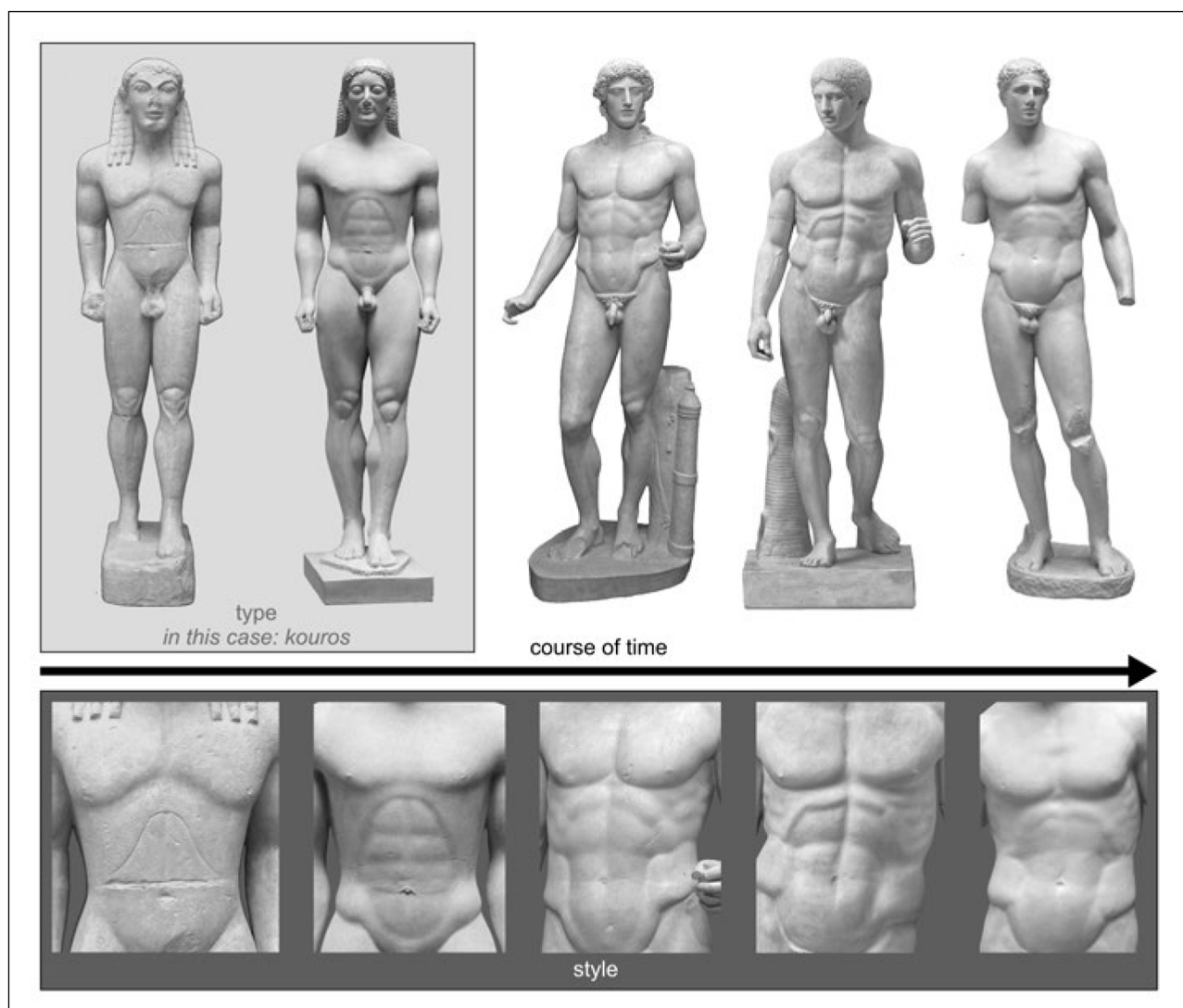


Figure 1 Exemplary illustration of type (top left) and style (bottom) using male statues of the Archaic and Classical periods as examples.

well as the principles by which these discrete forms are integrated to constitute a cohesive whole (Borbein 2000: 110–111, 121–122; von den Hoff 2019: 23–25).

Formal analysis therefore relies on identifying differences and similarities of form and recognizing recurring patterns. Traditionally, this is achieved through the visual comparison of sculptures, ideally original objects, though more often, given their global dispersion, via plaster casts or, most commonly, photographs. As such analyses are typically performed by individual researchers, the method, while fundamentally sound, remains susceptible to subjective interpretation. Another challenge associated with formal analysis concerns time: Given the vast number of extant sculptures from Greek and Roman antiquity – iDAI.objects Arachne,¹ the world’s most important archaeological object database primarily dedicated to ancient sculpture, currently lists more than 220,000 single objects, among which more than 100,000 sculptures – a comprehensive formal analysis by archaeologists may take weeks, months or even years, depending on the size and complexity of the sculptural group under investigation (for the formal analysis of the approximately 220 known portraits of Augustus within the publication series “Das römische Herrscherbild” of the German Archaeological Institute, Dietrich Boschung required about 2.5 years, even though he was able to build on extensive preliminary work (Boschung 1993a: XIII–XIV).

In consideration of these challenges, the present paper explores the application of deep learning algorithms based on Convolutional Neural Networks (CNNs) (Alzubaidi et al. 2021; Thakur et al. 2025) to the formal analysis of ancient sculptures. These methods enable an automatic comparison of images and the identification of shape similarities and differences – requiring significantly less time than manual analysis. This would result in a significant acceleration of the preliminary phase of any study of ancient sculptures. Consequently, archaeologists would be able to allocate more time to the scientifically demanding task of interpreting the results. And, last but not least, it would make formal analysis more objective. We therefore propose a CNN-based approach to automatically analyze images of ancient sculptures. This can detect and classify important structures and sort sculptures into groups.

Our test cases are the first four Roman emperors of the Julio-Claudian dynasty (Augustus, Tiberius, Caligula and Claudius). Not because we aim at identification but because the typology of those four emperors is already sufficiently well established (Augustus: (Boschung 1993a); Tiberius: (Hertel 2013); Caligula: (Boschung 1989); Claudius: (Boschung 1993b: 70–71)) so that the automatically computed results can be cross-checked with those reached by the traditional way. Furthermore, they offer an opportunity to address a severe challenge encountered when analyzing ancient sculptures using

computational methods: While certain groups – such as the approximately 950 surviving statues of Roman togati – are too large for a detailed analysis by humans within a reasonable timeframe, they are still modest in scale from the perspective of modern computational capabilities. And many sculptural groups are even (significantly) smaller: For instance, although approximately 220 portraits of the first Roman emperor, Augustus, have survived, they can be subdivided into distinct portrait types. Of these, only the *Prima Porta* type is attested in larger numbers, whereas all other types are represented by only a few examples (i.e. *Béziers* - *Spoletto* [2 replicas], *Lucus Feroniae* [3 replicas], *Alcudia* [29 replicas] and *Louvre MA 1280* [15 replicas in the main group, plus 11 and 4 in two sub groups]). Using intentionally small datasets of the first four Roman emperors as test cases therefore allows us to examine the performance of our proposed models in the face of limited data.

2 STATE OF THE ART

As noted in the introduction (Section 1), formal analysis of ancient sculptures continues to be conducted manually, as it has been for over a century. The method itself originated alongside the emergence of Classical Archaeology as a scientific discipline in the 18th century (Borbein 2000: 111–114). Despite considerable evolution over time, it remains primarily based on the visual comparison of individual sculptures by a single archaeologist, typically supported by photographic material.

In recent years, several attempts have been made to incorporate computational techniques into this analytical process. Most of these initiatives focus on the identification of Roman emperors, employing either 3D shape comparison (Lu et al. 2013; Langner 2021) or facial recognition algorithms (Ramesh et al. 2022; Lestriandoko et al. 2024; Heijnen et al. 2025). Both approaches prioritize facial features over hairstyle, thereby seeking to complement the traditional typological method, which relies predominantly on the analysis of coiffures – particularly the number and arrangement of locks framing the forehead (Fittschen 2015: esp. 62).² The facial recognition studies, in particular, aim to enable the identification of emperors based solely on facial features (Ramesh et al. 2022; Lestriandoko et al. 2024; Heijnen et al. 2025). This approach holds the potential to identify atypical or previously unattributed emperor portraits, especially in cases where conventional typological analysis remains inconclusive.

To date, only two projects have attempted a form of computer-assisted formal analysis of ancient sculptures that extends beyond typology. Both address the stylistic problem of master hand attribution in Greek sculpture and rely on 3D shape comparison techniques (Zhang

et al. 2013; Patay-Horváth 2014; Patay-Horváth & Christiansen 2017).

Furthermore, deep neural networks are employed for similar classification tasks in other archaeological contexts, such as the detection of facial features to assist in the identification of the Terracotta Warriors (Sheng 2024). Other examples include the use of deep learning to analyze cultural variations in the perception of facial expressions (Brooks et al. 2024), human pose estimation for artwork classification (Kutrzyński & Król 2024), and the handling of imbalanced datasets in the classification of cultural heritage images (Wisniamongkol et al. 2024). Machine learning methods have also been applied to ceramic classification tasks, as demonstrated in Portuguese case studies involving Lusitanian amphorae and faience (Santos et al. 2024).

Despite these developments, a broader and more systematic approach to algorithm-based formal analysis of ancient sculptures – one capable of conducting both typological and stylistic evaluations across a wide range of object types, including statues, busts, and reliefs and applicable to both small and large artifact groups – has yet to be developed.

However, the application of deep learning methods to archaeology faces several methodological challenges that complicate such an undertaking. Archaeological datasets are often comparatively small, which limits the effective training of large neural network architectures and increases the risk of overfitting. Transfer learning from models pre-trained on large-scale datasets such as ImageNet is therefore commonly employed to mitigate this limitation (Fiorucci et al. 2020). In addition, archaeological image collections frequently exhibit considerable variability in lighting conditions, viewing angles, documentation quality, and states of preservation, which may reduce model robustness and generalization across datasets. These challenges are also reflected in the datasets used in the present study, as the limited number of surviving sculptures and consistently documented images often prevents the creation of large training datasets in Classical Archaeology. The present work investigates whether meaningful results can be achieved under such conditions. In addition to classification performance, interpretability methods such as GradCAM enable the visualization of image regions relevant to the model's decisions. Comparing these regions with traditional features in archaeological analysis may offer new perspectives and encourage the wider availability of image datasets for future research.

As a first step towards such a goal we investigate two approaches in this paper: Image segmentation and classification. The latter is a first attempt at a typological analysis. And the former could in future be used to limit the classification task to selected regions of sculptures as e.g. the hair in portraits or the heads, the bodies or specific garments in statues. Although the two approaches are

interrelated, they are treated independently in this paper. Our primary aim is to assess the general applicability of deep learning techniques to such tasks within Classical Archaeology. The subsequent sections first describe the segmentation and classification methods employed, then present the results and discuss their implications.

3 METHODS

In order to evaluate the potential of deep learning methods for the analysis of ancient sculptures, the proposed approach is divided into two parts: segmentation and classification. The segmentation task of structural regions is addressed in Section 3.1. In Section 3.2, images of portrait heads are processed by a classification network to determine the depicted emperor or the corresponding portrait type, thereby enabling the categorization of the sculptures.

3.1 SEGMENTATION

The process of image segmentation partitions an image into multiple segments or regions of interest. In our test cases, those regions may represent facial features and other components of the head, such as the nose, mouth, ears, and hair within the given portrait images. In the case of 2D images, this is achieved by assigning a label to each pixel. While traditional image segmentation algorithms predominantly operate on some kind of uniformity criterion, deep learning based methods may exploit complex pattern recognition criteria for segmentation. The segmentation information can then be used for simple image region highlighting tasks, or for identifying image regions for further processing. Some of these methods contain the intermediate step of creating a binary mask for image regions of interest, where the pixels inside the region are assigned the value 1 and the remaining pixels a value of 0.

For our initial image segmentation, we process the images using a mask R-CNN deep learning model with a ResNet-50-FPN backbone (He et al. 2017; Lin et al. 2017). A mask R-CNN is a Region-based Convolutional Neural Network, a deep learning model for efficient object detection, which outputs binary image masks to identify the image regions associated with the detected objects. ResNet is a specific so-called residual neural network (He et al. 2016), a deep learning architecture where the neural network consists of usually multiple layers that learn residual functions to minimize the difference between observation and a predicted output. In this case, we use a ResNet serving as a feature extraction network that transforms the input image into hierarchical feature maps. The ResNet is embedded into a Feature Pyramid Network (FPN) (Lin et al. 2017), which is a neural network architecture optimized for detecting image features at different scales to allow for size-invariant object

detection. Mask R-CNN performs object localization and instance segmentation within a single inference pass. This allows the model to identify relevant regions and generate pixel-wise masks for individual structures without requiring a separate detection or cropping stage. Such an architecture is particularly suitable for the present task, as it enables the segmentation of multiple facial components (e.g. nose, eyes, ears, or hair) while preserving their spatial relationships (He et al. 2017). Alternative segmentation architectures such as fully convolutional networks (Long, Shelhamer & Darrell 2015) or U-Net models (Ronneberger, Fischer & Brox 2015) focus on dense pixel-wise prediction but do not explicitly include a region proposal stage. By combining detection and segmentation within a single framework, mask R-CNN enables the explicit localization and segmentation of structural image regions, which facilitates the analysis of facial components in portrait images.

Furthermore, we exploit the advantage that neural networks can be pre-trained using generic data, e.g. on training data for recognition of human facial features, and then be fine-tuned on a specific dataset to optimize for non-generic applications. This can reduce the amount of required application-specific training data significantly. Thus the mask R-CNN model with a ResNet-50-FPN backbone is initialized from pre-trained weights to obtain initial model parameters (Deng et al. 2009) and fine-tuned on the provided segmentation dataset (see Section 4.1) containing expert-labeled images of ancient sculptures.

The first step in the used R-CNN model is the identification of possible object candidates with the model's so-called region proposal network (RPN), which is followed by computing binary masks, class scores and

bounding boxes for the proposed regions. The individual class scores are then used to predict a class label for the regions.

On a formal level, for a given input image $\mathbf{I}_i \in \mathbb{R}^{W \times H \times C}$ with width, height and color $W, H, C \in \mathbb{N}$ this results in a set $\mathcal{S} = (\mathcal{B}, \mathcal{M}, \mathcal{L})$ for each input image. The bounding boxes ($\mathcal{B}$), masks ($\mathcal{M}$) and labels ($\mathcal{L}$) are predicted for each detected region e.g. the nose or the mouth. Each bounding box $\mathbf{b}_i \in \mathcal{B}$ is defined as a four-dimensional vector $\mathbf{b}_i = (x_1, y_1, x_2, y_2)$, where (x_1, y_1) and (x_2, y_2) denote the top-left and bottom-right coordinates of the box, respectively. Each corresponding mask $\mathbf{M}_i \in \mathcal{M}$ is a binary matrix of size $W \times H$ with values in 0, 1, indicating the pixel-wise segmentation of the i -th object. The label $\ell_i \in \mathcal{L}$ assigns the predicted class (e.g. “nose”, “mouth”, etc.) to the i -th region. Therefore, for each of the n detected regions, the model provides a tuple $(\mathbf{b}_i, \mathbf{M}_i, \ell_i)$ with $i = 1, \dots, n$. An example of a segmentation is depicted in Figure 2a.

3.2 CLASSIFICATION

The goal of the classification task is to assign a label to each input image (i.e. the name of the respective emperor or portrait type). For this, two separate pre-trained ResNet-18-models are used to classify the input images. Due to their robust feature representations and the availability of pre-trained weights, ResNet architectures are widely used as baseline models for image classification tasks (He et al. 2016; Alzubaidi et al. 2021). Recent transformer-based architectures have demonstrated strong performance on large-scale datasets, but often rely on substantially larger training corpora to reach their full potential (Dosovitskiy et al. 2021). Considering the limited size of the available

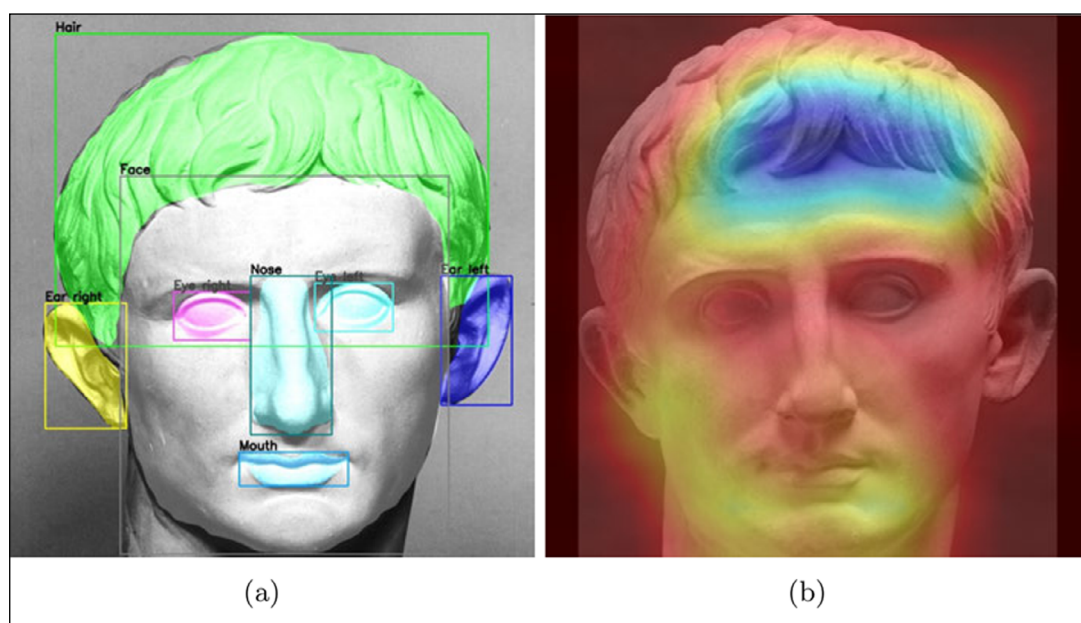


Figure 2 Comparative representation between (a) pixel-based segmentation and (b) GradCAM-generated visualization of regions significant for classification.

archaeological dataset, a convolutional architecture therefore provides a robust and well-established choice for the present classification task. To obtain the desired classifications in terms of emperor and portrait type, we trained one model for emperor classification and one model for portrait type classification. Additionally, to ensure better interpretability of the classification results and to enable comparison with traditional approaches, a visualization of the most relevant image regions is integrated.

Before the images are passed into the classification network, zero-padding is applied to ensure consistent and distortion-free image dimensions. Given an input image $\mathbf{I}_{in}$ with width W , height H , and number of color channels C , the image is padded such that the longer side determines the target square size while maintaining the original aspect ratio. Let $S = \max(W, H)$ denote the side length of the resulting square image. The image is then padded with zeros on the shorter side to form a square image $\mathbf{I}_{pad} \in \mathbb{R}^{S \times S \times C}$. Subsequently, the padded image $\mathbf{I}_{pad}$ is resized to a fixed resolution $\mathbf{I}_{resized} \in \mathbb{N}^{R \times R \times C}$, where $R \in \mathbb{N}$ is the target input size of the classification network. The padded and resized image $\mathbf{I}_{resized}$ is then fed into the classification network, which outputs a predicted class label ℓ_i corresponding to the emperor or the portrait type.

An inherent property of most deep learning models is the fact that the decision process leading to a certain classification, for example which image regions the classification decision was based on, cannot be intuitively understood by humans without aid. This issue can be addressed by methods from the field of explainable artificial intelligence (AI), which aim to make AI predictions comprehensible to humans (Yang et al. 2023). For comparing established, traditional classification criteria, like the shape of hair curls, with the classification criteria from our trained models, we employ a Gradient-weighted Class Activation Mapping (GradCAM) (Selvaraju et al. 2017) approach. GradCAM can compute a heatmap for highlighting the image regions that contributed the most to the final class prediction (see Figure 2b). While GradCAM provides only an approximate visualization of the internal decision-making process of the network, it offers the opportunity to compare the regions considered relevant by the model with those traditionally emphasized by archaeologists. By highlighting similarities and differences between human visual analysis and algorithmic feature extraction, the approach may help identify overlooked diagnostic features or compare the relative importance of established ones. This allows evaluating the decision making process with regard to the plausibility of the classification features used.

However, GradCAM heatmaps do not represent a direct explanation of the network's reasoning (Selvaraju et al. 2017). They highlight regions that contribute most strongly to the prediction according to gradient-based

approximations. For this reason, GradCAM visualizations in this study are primarily used as a qualitative tool to facilitate comparison between computational and traditional visual analysis, rather than as a conclusive explanation of the model's internal reasoning.

4 RESULTS AND DISCUSSION

In this section, the results obtained from the dataset in combination with the introduced segmentation and classification methods (see Section 3) are presented and contextualized. First, we describe the dataset (Section 4.1) as well as the experimental setup and relevant parameters used in the experiments (Section 4.2). Then, we report the outcomes of the segmentation procedure (Section 4.3), followed by the classification results (Section 4.4), which are further divided into emperor recognition and portrait type classification.

4.1 DATASET

In order to evaluate the models' performance on segmentation and classification tasks, different datasets and annotations are required.

Our presented segmentation model (Section 3.1) assigns a label to each image pixel. Therefore, a labeled dataset with annotations for all relevant regions is necessary. For this purpose, we created a dataset containing 32 different images of portraits of the emperor Augustus and manually annotated each image to delineate specific facial regions, including the mouth, nose, eyes, ears, and other key features of the head. Furthermore, in order to analyze the segmentation model's generalization capability, an additional dataset consisting of 11 images of other Roman emperors (Tiberius, Caligula, and Claudius) was annotated. These images were not included in the training process and were used exclusively for evaluation. Our proposed classification models (Section 3.2) predict a single label for each input image. In order to compare the predicted labels with the ground truth, each image in the dataset must be provided with a corresponding ground truth label. For the purpose of classifying different Roman emperors a labeled dataset with portrait images of the first four emperors (Augustus, Tiberius, Caligula and Claudius) was utilized. Each emperor is represented by 24 images, resulting in a balanced dataset of 96 images. The images were selected based on established attributions in the scholarly literature to ensure both authenticity and consistency in labeling.

In addition, a separate labeled dataset was employed for the classification of portrait types, using the example of Augustus and differentiating between the type *Louvre MA 1280* and the type *Prima Porta*. Each portrait type is represented by 17 images, resulting in a total of 34 images.

As stated in the introduction (Section 1), the datasets employed for all tasks were intentionally kept small in order to evaluate the models' performances under conditions of limited data.

4.2 EXPERIMENTAL SETUP

Because the datasets are comparatively small (Section 4.1) we used stratified five-fold cross-validation to measure the performance of the deployed models (Bishop & Nasrabadi 2006). Stratified five-fold cross-validation partitions the dataset into five folds while preserving the class distribution across folds (Bishop & Nasrabadi 2006). In each fold, approximately 80% of the images were used as the training data and 20% as the test data. Within the training images of each fold, 20% of the images were reserved as a validation subset for model selection. Exact fold sizes varied slightly because the dataset size was not divisible by five. Each fold is used once as the test set, while the remaining folds are used for training. Model fitting was performed on the training subset, checkpoint selection on the validation subset, and final performance evaluation on the test fold. During training, data augmentation techniques such as small rotations and affine transformations were applied in the data loading process to improve generalization.

For the segmentation task, a mask R-CNN model with a ResNet-50-FPN backbone was employed. The model was initialized from pretrained weights and fine-tuned on the annotated segmentation dataset. The model was trained using an Adam optimizer with a learning rate of 10^{-4} and a weight decay of 10^{-4} (Kingma & Ba 2017). Training was performed for 50 epochs. The batch size was set to 4 for training and 1 for evaluation. For each fold, the model achieving the best performance on the validation subset was selected and subsequently evaluated on the corresponding test fold. During inference, predicted instance masks were binarized using a fixed threshold of 0.5.

The models for the classification tasks were initialized with *ImageNet* pretrained weights and optimized using the Adam optimizer (Kingma & Ba 2017). We conducted a small predefined grid search within each fold over the learning rate and weight decay parameters. The tested configurations were: (1) learning rate 1×10^{-4} with weight decay 0, (2) learning rate 1×10^{-4} with weight decay 1×10^{-4} , and (3) learning rate 3×10^{-4} with weight decay 1×10^{-4} . Model selection was based on validation

macro-F1. Early stopping with a patience of 15 epochs and a minimum improvement threshold of 10^{-4} was applied.

4.3 SEGMENTATION

Our proposed segmentation approach (Section 3.1) divides images into different regions of the head, such as the nose, mouth, and other facial areas. For an objective evaluation of the segmentation performance, the Intersection over Union (IoU) score is used to report the results (Everingham et al. 2010). For each class and image, IoU was computed if the class was present in either the ground truth or the prediction. Cases where the class was missing in both were excluded from evaluation. The IoU is a common metric for evaluating segmentation performance and is defined as the ratio of the overlap area between predicted and ground truth segmentation to their combined area, ranging from 0 (no overlap) to 1 (perfect overlap). In general, IoU values above 0.5 are classified as acceptable (Everingham et al. 2010).

For each fold, the average IoU score across all classes (nose, hair, etc.) and all images is calculated on the resulting test dataset. Mean IoU values and standard deviations (Std) for each class, indicating the variability, are presented in Table 1. Overall the averaged IoU score is 0.8496 with a standard deviation of ± 0.0301 , indicating low variability across folds. Furthermore, IoU scores are presented in Figure 3. The boxplots (box-and-whisker plots) summarize the distribution of IoU values across folds, where the median is shown as a central line, the box represents the interquartile range, whilst the 'whiskers' represent the spread of the data (excluding outliers), and any outliers (circles in the plot) are marked separately. The blue boxplots represent the distribution of IoU values across all classes. Additionally, two selected classes, hair (green) and left eye (red), are shown to illustrate the variation in segmentation performance between individual classes. Overall, the results show consistently high IoU scores with moderate variability across folds, while certain classes show slightly larger variation. Outliers, indicated by circles, are observed for the hair class. Additionally, outliers in the overall distribution are most likely caused by missing ear segmentations. While the quantitative metrics provide an overall indication of segmentation quality, visual inspection of selected examples helps to illustrate these values. A visual interpretation of the IoU values can be made by comparing the differences between the images

METRIC	HAIR	L. EAR	L. EYE	R. EYE	R. EAR	FACE	MOUTH	NOSE
IoU	0.845	0.761	0.908	0.892	0.779	0.941	0.791	0.878
Std	0.040	0.102	0.013	0.016	0.094	0.011	0.068	0.044

Table 1 Segmentation performance measured using Intersection over Union. Mean IoU values and standard deviations across the five evaluation folds are reported for each class.

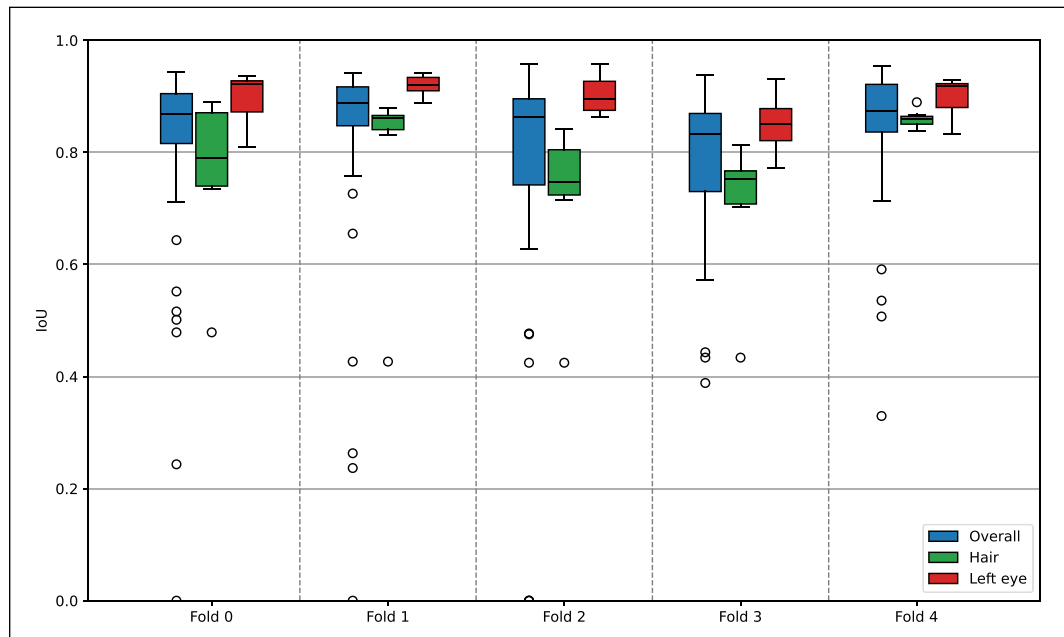


Figure 3 Distribution of the IoU values across the five test folds for all classes combined (blue), hair (green) and left eye (red). The folds are separated by dashed lines.

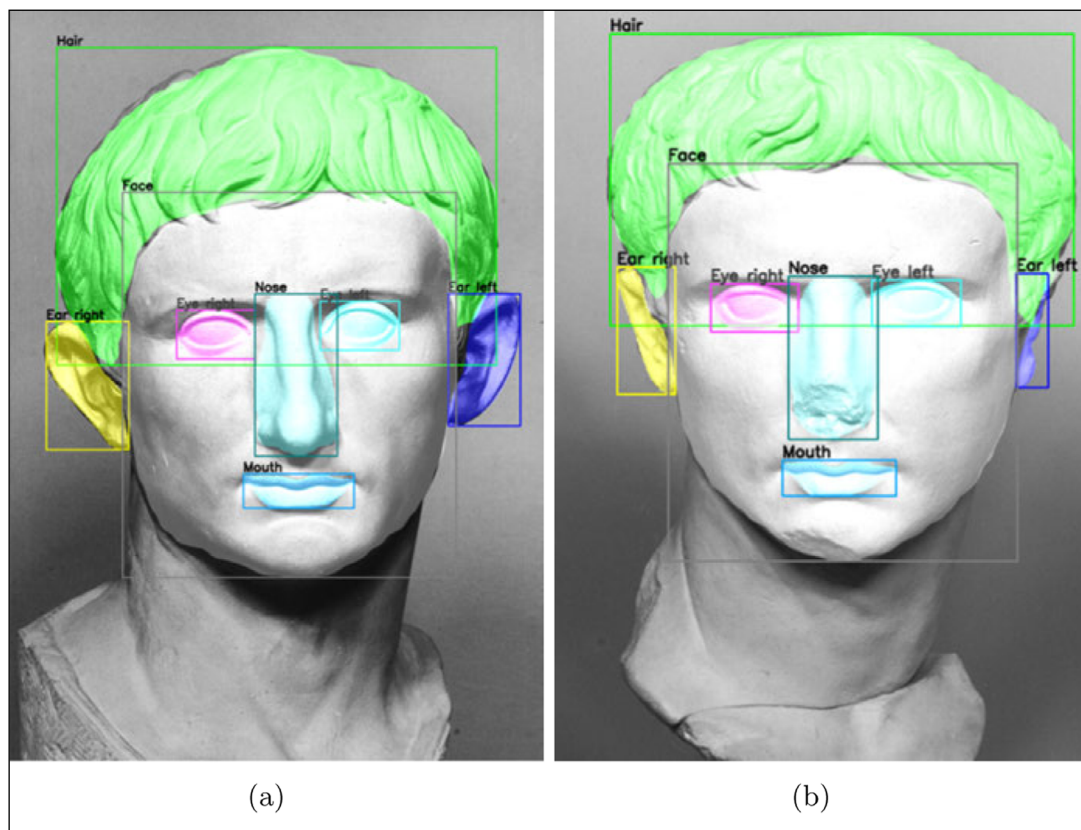


Figure 4 Two segmentation examples of portraits of Augustus with IoU of 0.912 (a) and 0.885 (b), computed from a model trained on the same fold.

shown in Figure 4. Two full frontal views of portraits of Augustus were segmented with a model derived from the same training fold, resulting in IoU scores of 0.912 for Figure 4a and 0.885 for Figure 4b. The different regions predicted by the model are visualized for each class (e.g. mouth, hair, nose) using overlapping colors with semi-transparent alpha values. It can be observed that in

both images, the facial mask slightly overlaps with the hair region. Furthermore, the motif of locks falling across the forehead, which forms an interlocking pattern, is simplified by the model: parts of the brow are classified as hair, while the tips of the locks remain unsegmented. Additional unsegmented areas are observed in the hair around the ears and along the contours of both heads,

with these effects being more pronounced in [Figure 4b](#) than in [Figure 4a](#). This is consistent with the respective IoU-scores and the results illustrated in [Figure 3](#), where larger hair regions with complicated edges tend to show lower IoU values (green boxplot). In contrast, smaller and more distinct regions, such as the left eye, are annotated more accurately, as can be seen visually in [Figure 4](#) (light blue) and quantitatively in [Figure 3](#) through higher IoU scores (red boxplot).

From an archaeological perspective, however, the results obtained with limited training data are likely already sufficient for comparative analysis, as minor overlaps between masks or small unsegmented regions of hair at the temples and along the contours of the head in a full-frontal view are unlikely to significantly affect the overall outcome. Naturally this will have to be verified in future experiments. And it will also have to be tested whether the performance of the model can be improved by using more training data. In view of the limited size of some archaeological groups (as mentioned in the introduction Section 1) it will also have to be determined whether there exists a threshold of applicability and a point at which more training data does not lead to significantly better results.

As the dataset mainly consists of frontal and near-frontal portrait views, the model exhibits reduced performance when encountering three-quarter and side views, which is reflected in lower IoU scores as presented in [Figure 5](#). Especially, in this case, the regions around the ears are missegmented, mainly due to incorrect hair pixel labels, resulting in overlapping and missing regions of

hair, and face. This effect is indicated by the overlapping color regions, which can serve as an explanation for the outliers observed in [Figure 3](#) (green boxplot). Moreover, part of the left side of the face is missing from the predicted facial region, and the hair and face masks show substantial overlap at both temples. Finally, the left ear—although not visible in the image due to the pronounced turn of the head—is apparently inferred by the model and is instead misclassified as part of the garment draped over the left shoulder. As full-frontal views – despite their general prevalence in Classical Archaeology – are not available for every ancient sculpture, this limitation of the current model will need to be addressed through training future versions on images from multiple viewpoints. This should also include whole statues as well as reliefs, which exhibit properties distinct from those of free-standing sculptures.

Furthermore, the model was evaluated for its ability to segment images of other Roman emperors, despite having been trained exclusively on portraits of Augustus. To this end, additional images of Tiberius, Caligula and Claudius were annotated and fed into the model for comparison. Some resulting segmentations are shown in [Figure 6](#) for a portrait of Caligula ([Figure 6a](#)) and in ([Figure 6b](#)) for another one of Tiberius. They exhibit the same strengths and weaknesses as those witnessed for the portraits of Augustus. This suggests that a model trained to segment human forms can be expected to perform consistently across different representations of the human figure, regardless of the specific individuals included in the training data.

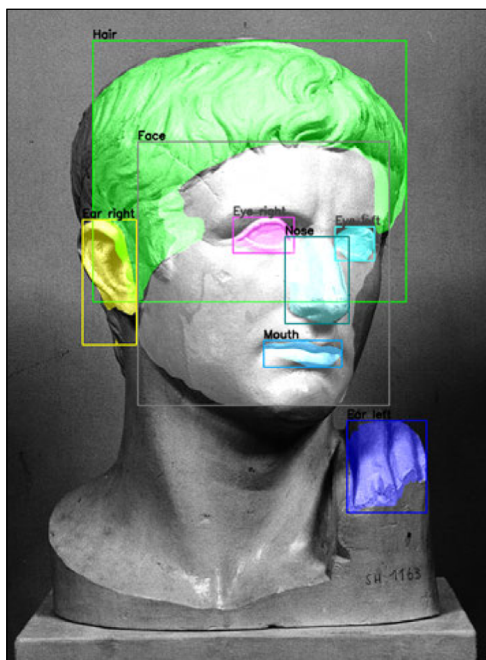


Figure 5 A portrait of Augustus in three-quarter view where the segmentation becomes worse, missing parts of the hair on the right side, failing to differentiate correctly between hair and face at the temples and labeling the garment over the left shoulder incorrectly as the left ear.

4.4 CLASSIFICATION

For the classification task we proposed a model (Section 3.2) that is capable of assigning a single label to each input image. Depending on the task, the results are presented in Section 4.4.1 for emperor recognition and in Section 4.4.2 for portrait type estimation. Both tasks are based on the same model but were trained and tested on different datasets, as presented in Section 4.1.

4.4.1 Emperor

The results for identifying the depicted emperor in a given input image, based on the model from Section 3.2, are presented in the following. The classification approach was applied to predict the four emperors: Augustus, Tiberius, Caligula and Claudius. For quantitative evaluation of the classification tasks, several standard performance metrics were used, including accuracy, precision, recall, and F1-score ([Sokolova & Lapalme 2009](#)). Accuracy measures the amount of correctly classified images relative to the total number of samples. Precision quantifies the number of correctly predicted instances among all samples assigned to a given class. Recall measures the number of correctly identified instances relative to all ground truth samples of that class. The

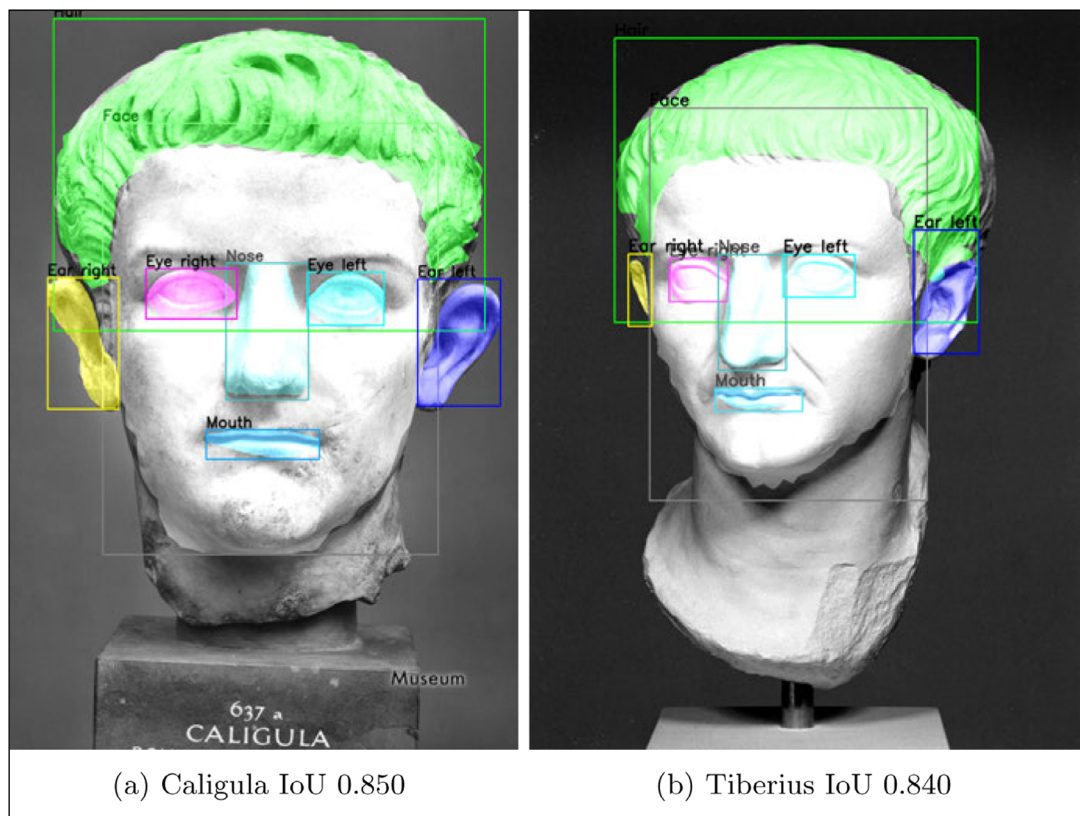


Figure 6 Segmentation results for the two emperors Caligula **(a)** and Tiberius **(b)**. The test images differ from the training data, which contains only portraits of Augustus.

F1-score represents the harmonic mean of precision and recall and combines both aspects in a single metric. In this work, macro-averaged precision, recall, and F1-score are reported. Macro-averaging computes each metric independently for every class and then averages the results across classes (Sokolova & Lapalme 2009). In addition, to provide an overview of the model's performance, a normalized confusion matrix (Figure 7) illustrates the correspondence between predicted and annotated image labels across the stratified five-fold cross-validation.

The normalized confusion matrix depicts differences in classification performance across the four emperors. Augustus achieves the highest recognition rate with approximately 88% of images correctly classified, followed by Claudius with about 83%. Caligula is correctly identified in roughly 75% of the cases. The lowest recognition rate is observed for Tiberius, with approximately 67% of images correctly classified. Misclassifications of Tiberius frequently occur as Caligula, indicating a strong visual similarity between these portraits from the perspective of the model.

Following the visual overview provided by the confusion matrix, the quantitative evaluation metrics are summarized in Table 2. The table reports the overall classification performance across the stratified five-fold cross-validation. In addition, Table 3 presents the

per-class precision, recall, and F1-score values for each emperor.

Overall, the classification results indicate that the model is able to distinguish between the four emperors with a reasonable level of accuracy. The averaged metrics show balanced performance across the classes, with a macro F1-score of approximately 0.78. The per-class results further reveal differences in classification difficulty. Augustus and Claudius achieve comparatively higher scores, while Tiberius shows the lowest precision, recall, and F1-score values. This observation is consistent with the confusion matrix, where portraits of Tiberius are more frequently misclassified as Caligula or Claudius.

This performance of the model despite limited training data exceeded expectations. As with the segmentation task, future research will need to determine whether the results improve with additional training data, particularly in the case of Tiberius. Two of the misclassifications – one instance of Tiberius classified as Caligula and one of Caligula classified as Tiberius – are likely influenced by the fact that the respective images depict the portraits from viewing angles that are atypical within the dataset (namely, a pronounced low-angle view and a lateral view), whereas the model was trained predominantly with full-frontal images. As in the case of the segmentation task, this limitation of the current model will need to be addressed by training future versions on a broader range of viewing angles.

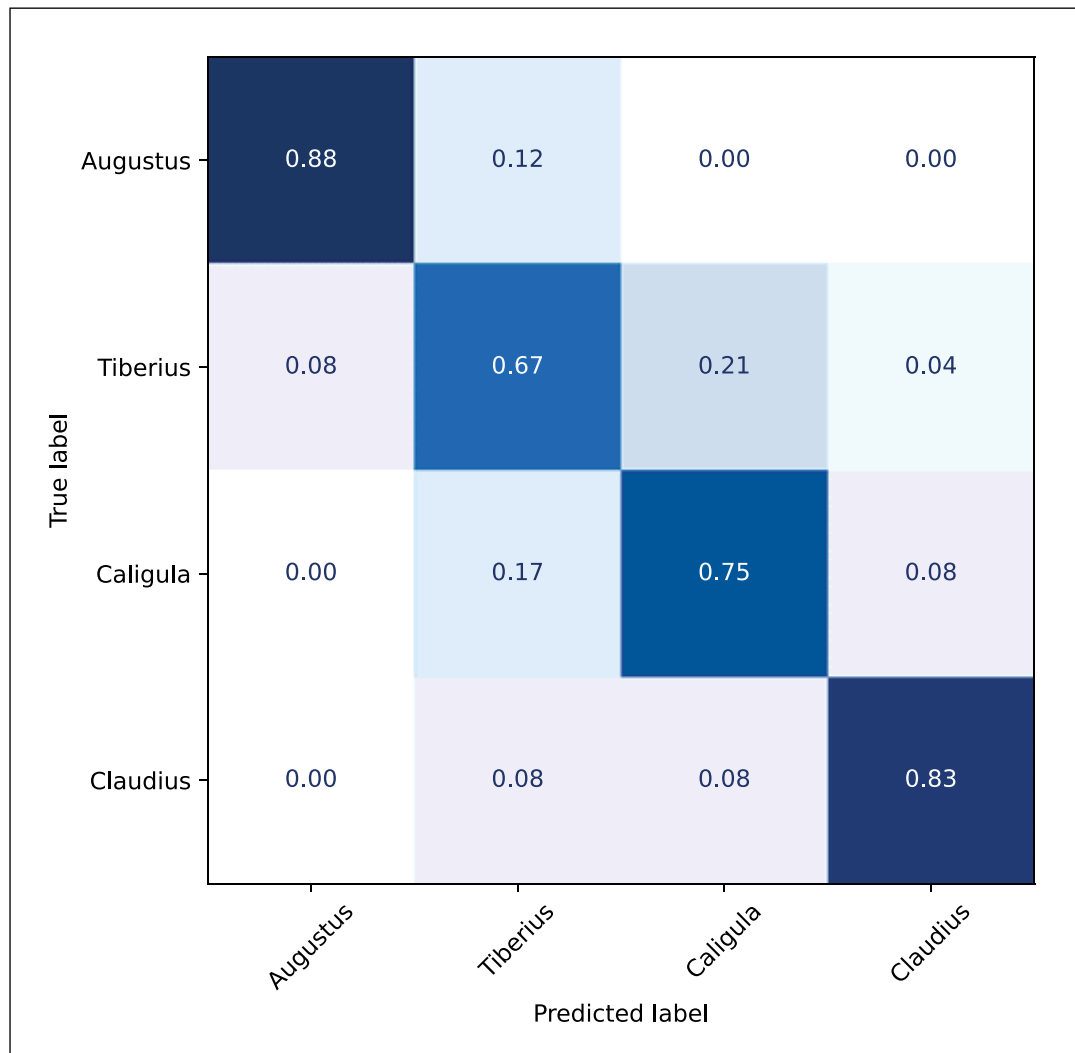


Figure 7 The normalized confusion matrix for the emperor classification task. Since the test predictions were aggregated across the five folds, each emperor class contains 24 images in total. Therefore, a single misclassification corresponds to approximately 0.04 in the normalized matrix.

METRIC	RESULT
Accuracy	0.7821 ± 0.1046
Macro Precision	0.8249 ± 0.0737
Macro Recall	0.7850 ± 0.1066
Macro F1	0.7807 ± 0.1065

Table 2 Emperor classification results obtained by stratified five-fold cross-validation. Values are reported as mean ± standard deviation across folds.

Furthermore, a GradCAM approach (see Section 3.2) was used to indicate which image regions contributed to the model's prediction. A comparison between a correct and an incorrect emperor classification is shown in Figure 8 where Augustus is recognized correctly in Figure 8a and incorrectly classified as Tiberius in Figure 8b. The blue areas highlight the most decisive regions for the classification process, with decreasing relevance indicated by yellow, and least relevant or negligible regions shown in red.

According to these results, the hair region appears to have been particularly relevant for the neural network in determining the predicted class. In 73% of the test cases, it appears to have been the primary deciding factor, leading to a correct classification in 55 cases and to an erroneous one in only 15 cases. This may be cautiously interpreted as corroborating the traditional archaeological methods that scholars have employed for decades to identify the portraits of Roman emperors, especially given that the model was not explicitly trained to prioritize specific features of the portraits fed into it. On the other hand, the observation that, among the correctly identified emperors, 19 attributions appear to rest predominantly on facial features, aligns well with recent scholarship emphasizing the importance of the face in imperial portraiture (Langner 2021; Ramesh et al. 2022; Lestriandoko et al. 2024; Heijnen et al. 2025). In particular, the neural network seems to have relied strongly on the region around the eyes as well as the mouth for attribution based on facial features. The finding that Tiberius is apparently

EMPEROR	PRECISION	RECALL	F1-SCORE
Augustus	0.9333 $\pm$ 0.0913	0.8700 $\pm$ 0.1857	0.8851 $\pm$ 0.0914
Tiberius	0.6933 $\pm$ 0.2186	0.6700 $\pm$ 0.1718	0.6524 $\pm$ 0.1286
Caligula	0.7800 $\pm$ 0.1789	0.7600 $\pm$ 0.2191	0.7351 $\pm$ 0.1353
Claudius	0.8929 $\pm$ 0.1473	0.8400 $\pm$ 0.2191	0.8500 $\pm$ 0.1491

Table 3 Per-class performance for emperor classification. Values are reported as mean $\pm$ standard deviation across folds.

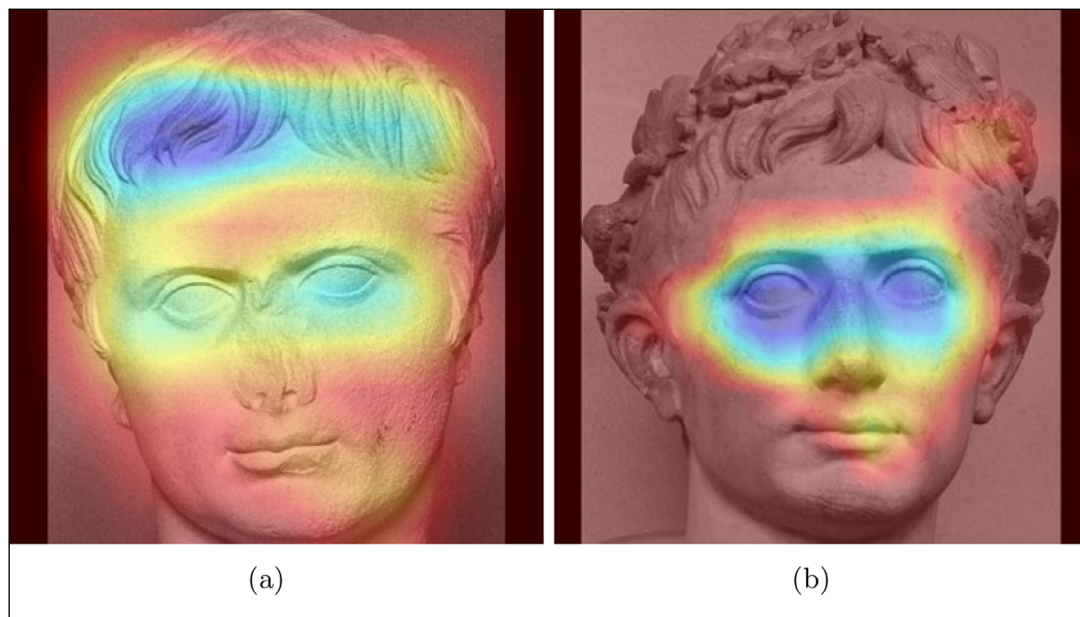


Figure 8 In (a) a portrait of Augustus is classified correctly as Augustus whereas in (b) another portrait of Augustus is erroneously classified as Tiberius.

classified correctly primarily on the basis of his facial features (and incorrectly mainly on the basis of his coiffure) is noteworthy; however, this result requires validation on a larger dataset before any interpretation can be offered. The results of 28 representative predictions (out of 96) are illustrated in an overview collage (see [Figure 9](#)), highlighting the important regions for each test image identified by the proposed GradCAM approach.

4.4.2 Portrait type

Analogous to the emperor classification presented in Section 4.4.1, the following section shows the results of the portrait type classification. For the two selected portrait types of Augustus (*Louvre MA 1280* and *Prima Porta*), predictions were generated using the stratified cross-validation. The normalized confusion matrix shown in [Figure 10](#) is computed from the aggregated predictions across all cross-validation folds.

Furthermore, the overall results for the metrics are presented in [Table 4](#) and the per class results are shown in [Table 5](#). The reported values represent the mean and standard deviation across the five cross-validation folds.

The quantitative results for the type classification indicate a consistent classification performance for

the two types. Across the five cross-validation folds, the average accuracy is 0.7952 and the macro F1-score is 0.7867. The per-class metrics show a balanced performance for both types. With a recall of 0.8333 the *Prima Porta* type achieves a slightly higher value compared to the *Louvre MA 1280* type (0.7833), while precision values are comparable for both classes. A noticeable limitation is the relatively large standard deviation for all metrics, which reflects the limited size of the dataset containing only 34 images in total. Since each test fold of the stratified five-fold cross-validation contains only a small number of samples, individual misclassifications can have a greater impact on the metrics. Nevertheless, the overall results still demonstrate the model's capability to capture type specific visual characteristics of the two types and to discriminate between them. The relatively balanced per-class metrics further suggest that the model does not strongly favor one class over the other.

To obtain an approximate indication of which image regions influenced the model's predictions in order to compare them to traditional features, Grad-CAM visualizations were generated for representative test images from the cross-validation folds ([Figure 11](#)). The figure shows example images, derived from several test folds, of the two portrait types (*Louvre MA 1280* and



Figure 9 Overview of GradCAM visualizations for representative emperor classification test images. The emperors are arranged in columns, starting with Augustus, followed by Tiberius, Caligula, and Claudius. Green borders indicate correct classifications, red borders indicate incorrect classifications.

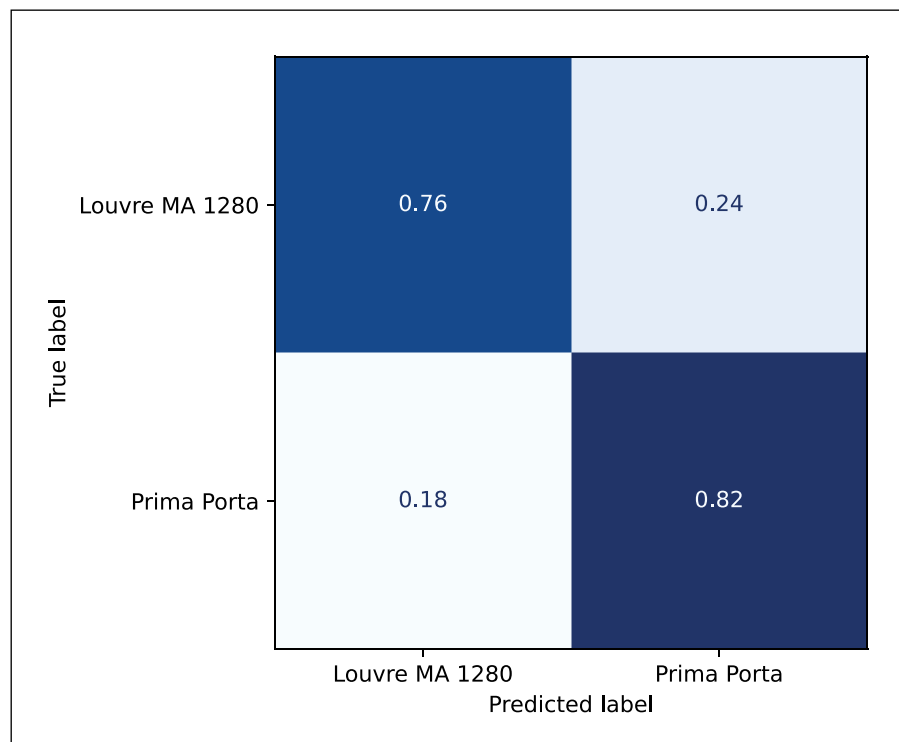


Figure 10 The normalized confusion matrix for the type classification task. Since the test predictions were aggregated across the five folds, each class contains 17 images in total. Therefore, a single misclassification corresponds to approximately 0.06 in the normalized matrix.

METRIC	RESULT
Accuracy	0.7952 ± 0.2153
Macro Precision	0.8417 ± 0.1778
Macro Recall	0.8083 ± 0.1990
Macro F1	0.7867 ± 0.2254

Table 4 Portrait type classification results obtained by stratified five-fold cross-validation. Values are reported as mean ± standard deviation across folds.

TYPUS	PRECISION	RECALL	F1-SCORE
Louvre MA 1280	0.8500 ± 0.2236	0.7833 ± 0.3312	0.7657 ± 0.2691
Prima Porta	0.8333 ± 0.2357	0.8333 ± 0.2357	0.8076 ± 0.1935

Table 5 Per-class performance for portrait-type classification. Values are reported as mean ± standard deviation across folds.

Prima Porta), including both correctly and incorrectly classified cases.

The GradCAM visualizations reveal patterns that are essentially identical to those identified in the emperor-classification task described in Section 4.4: The hair region again emerges as the dominant cue for the model's decisions and accounts for the majority of correct attributions (22 of 36 cases). The facial area – particularly the region surrounding the mouth – also

contributes to the predictions. However, the model assigns surprisingly little importance to what is, from a human perspective, the most salient characteristic of Augustan portraiture: the prominent fork-and-pincer-motif above the forehead in the *Prima Porta* type. Unlike in the preceding emperor-classification task (Figure 9), this feature is rarely highlighted in its entirety in dark blue during portrait type classification, suggesting that the neural network does not rely on it as a primary discriminative cue. Together with the markedly higher standard deviation observed for this classification task, this finding indicates that the training set is likely too limited to enable the model to capture the typological distinctions necessary to differentiate between the two classes. One possible mitigation strategy, warranting further investigation, would be to restrict the comparison to a previously segmented region such as the hair (see above).

5 CONCLUSION

The two approaches chosen here proved to be highly promising. The segmentation performed well despite the limited amount of training data. Its successful application to the emperors Tiberius and Caligula, although the model was trained exclusively on images of Augustus, highlights its potential to overcome data scarcity and reduce annotation effort. Further research will be required to determine whether the results can be improved through the inclusion of additional training

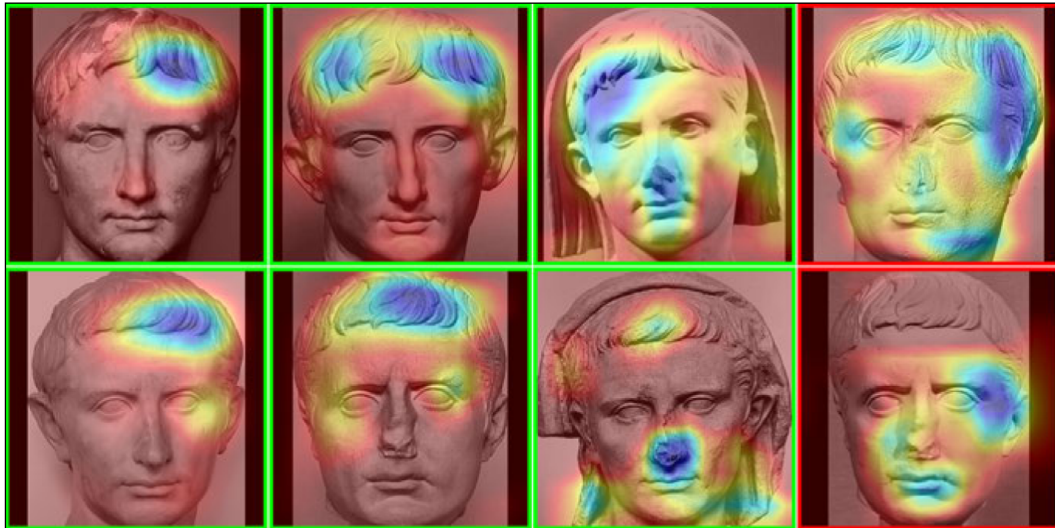


Figure 11 The example images were generated using the presented GradCAM approach on different test fold samples and highlight the regions that are decisive for the classification. The types *Prima Porta* and *Louvre MA 1280* are arranged in rows, with *Prima Porta* on top and *Louvre MA 1280* below. Green borders indicate correct classifications (first three images in each row), red borders indicate incorrect classifications (last image in each row).

data and whether a point exists at which increasing the dataset no longer leads to significant performance gains. Moreover, the model must be trained to account for 'archaeological reality', including varying viewing angles and different states of preservation. It will also be necessary to investigate whether the segmentation approach can be extended beyond the human face to bodies and their limbs, which may appear in diverse poses and either with or without garments. Nevertheless, if future training data encompass the full range of characteristics typical of ancient sculpture, the model should ultimately be capable of accurately segmenting any sculpture presented to it, even if that particular group was not included in the training data.

In addition, the performance of the classification task is encouraging. The results indicate that the model relied primarily on hairstyle, while also incorporating facial features for identification, despite not being explicitly instructed to do so. This behaviour closely mirrors established archaeological practice, in which hairstyle has traditionally served as the principal diagnostic criterion, while more recent scholarship has increasingly emphasized the importance of facial characteristics. This emergent convergence between computational behaviour and disciplinary methodology suggests that AI-based approaches hold considerable promise for typological classification. Such methods may not only substantially accelerate the analytical process but also enhance methodological consistency through the application of standardized decision criteria. Future research will need to assess whether the model's performance can be further improved, both with regard to portrait-type classification specifically and across a broader range of emperors. It will also be important to assess the reliability of classifications involving portraits in which the hair is no longer arranged in clearly defined

locks, as became fashionable in Roman portraiture from the mid-2nd century CE onward. In parallel with the segmentation approach, further investigation is required to determine whether and how the classification framework can be extended beyond the head to include – or even focus exclusively on – posture, garments, and other sculptural features.

NOTES

- 1 <https://arachne.dainst.org>, last accessed 03/06/2026.
- 2 Cf. also (Lettinga 2022) who focuses on the hair and uses a variety of machine learning methods.

SUPPLEMENTARY DATA

The code for programs and data used in the paper can be accessed on <https://graphics.cs.tu-dortmund.de/data/jaa>.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

All authors contributed equally.

LIST OF ILLUSTRATIONS

- [Figure 1](#): Collage of different photographs of the Museum für Abgüsse Klassischer Bildwerke München.

- [Figure 2](#): Museum für Abgüsse Klassischer Bildwerke München (a: Heide Glöckler, b: Roy Hessing).
- [Figure 4](#): Museum für Abgüsse Klassischer Bildwerke München.
- [Figure 5](#): Museum für Abgüsse Klassischer Bildwerke München (Roy Hessing).
- [Figure 6a](#): <https://arachne.dainst.org/entity/6807571>, Last accessed 07/24/2025 (Ny Carlsberg Glyptotek).
- [Figure 6b](#): Museum für Abgüsse Klassischer Bildwerke München.
- [Figure 8](#): Museum für Abgüsse Klassischer Bildwerke München (Roy Hessing).
- [Figure 9](#): Collage of different photographs of the Museum für Abgüsse Klassischer Bildwerke München and <https://arachne.dainst.org>, Last accessed 07/24/2025.
- [Figure 11](#): Collage of different photographs of the Museum für Abgüsse Klassischer Bildwerke München and <https://arachne.dainst.org>, Last accessed 07/24/2025.

AUTHOR AFFILIATIONS

Birgit Bergmann  orcid.org/0000-0002-9015-4550
Department of Archaeologies (Classical Archaeology),
University of Freiburg, 79085 Freiburg, Germany

Daniel Matuszczyk  orcid.org/0000-0001-6403-1488
Department of Computer Science VII, Technical University
Dortmund, 44227 Dortmund, Germany

Denis Fisseler  orcid.org/0000-0002-0796-9442
Department of Computer Science VII, Technical University
Dortmund, 44227 Dortmund, Germany

Thoralf Schröder
Institute of Archaeology (Classical Archaeology), University of
Cologne, 50937 Cologne, Germany

Nele Schröder-Griebel  orcid.org/0009-0005-4599-2650
Museum of Plaster Casts of Ancient Art, 80333 Munich,
Germany

Frank Weichert  orcid.org/0000-0002-2530-8197
Department of Computer Science VII, Technical University
Dortmund, 44227 Dortmund, Germany

REFERENCES

- Alzubaidi, L, Zhang, J, Humaidi, AJ, Al-Dujaili, A, Duan, Y, Al-Shamma, O, Santamaría, J, Fadhel, MA, Al-Amidie, M and Farhan, L.** 2021. 'Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions'. *Journal of Big Data*, 8: 1–74. DOI: <https://doi.org/10.1186/s40537-021-00444-8>
- Bishop, CM and Nasrabadi, NM.** 2006. *Pattern Recognition and Machine Learning*. Vol. 4. 4. Springer.
- Borbein, AH.** 2000. 'Formanalyse'. In: Borbein, AH, Hölscher, T and Zanker, P (Eds.) *Klassische Archäologie. Eine Einführung*. Berlin: Dietrich Reimer, pp. 109–128.
- Boschung, D.** 1989. *Die Bildnisse des Caligula*. Das römische Herrscherbild I 4. Berlin: Gebr. Mann.
- Boschung, D.** 1993a. *Die Bildnisse des Augustus*. Das römische Herrscherbild I 2. Berlin: Gebr. Mann.
- Boschung, D.** 1993b. 'Die Bildnistypen der iulisch-claudischen Kaiserfamilie: ein kritischer Forschungsbericht'. *Journal of Roman Archaeology*, 6: 39–79. DOI: <https://doi.org/10.1017/S1047759400011466>
- Brooks, JA, Kim, L, Opara, M, Keltner, D, Fang, X, Monroy, M, Corona, R, Tzirakis, P, Baird, A, Metrick, J, others.** 2024. 'Deep Learning Reveals What Facial Expressions Mean to People in Different Cultures'. *iScience*, 27(3). DOI: <https://doi.org/10.1016/j.isci.2024.109175>
- Deng, J, Dong, W, Socher, R, Li, L-J, Li, K and Fei-Fei, L.** 2009. 'ImageNet: A Large-Scale Hierarchical Image Database'. In: *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 248–255. DOI: <https://doi.org/10.1109/cvpr.2009.5206848>
- Dosovitskiy, A, Beyer, L, Kolesnikov, A, Weissenborn, D, Zhai, X, Unterthiner, T, Dehghani, M, Minderer, M, Heigold, G, Gelly, S, Uszkoreit, J and Houshy, N.** 2021. 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale'. arXiv: 2010.11929 [cs.CV]. <https://arxiv.org/abs/2010.11929>
- Everingham, M, Van Gool, L, Williams, CKI, Winn, J and Zisserman, A.** 2010. 'The Pascal Visual Object Classes (VOC) Challenge'. *International Journal of Computer Vision*, 88(2): 303–338. DOI: <https://doi.org/10.1007/s11263-009-0275-4>
- Fiorucci, M, Khoroshiltseva, M, Pontil, M, Travaglia, A, Del Bue, AI and James, S.** 2020. 'Machine Learning for Cultural Heritage: A Survey'. *Pattern Recognition Letters*, 133: 102–108. DOI: <https://doi.org/10.1016/j.patrec.2020.02.017>
- Fittschen, K.** 2015. 'Methodological Approaches to the Dating and Identification of Roman Portraits'. In: Borg, BE (Ed.) *A Companion to Roman Art*. Blackwell Companions to the Ancient World. Chichester: John Wiley & Sons, Ltd, pp. 52–70. DOI: <https://doi.org/10.1002/9781118886205>
- He, K, Gkioxari, G, Dollár, P and Girshick, R.** 2017. 'Mask R-CNN'. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969. DOI: <https://doi.org/10.1109/ICCV.2017.322>
- He, K, Zhang, X, Ren, S and Sun, J.** 2016. 'Deep Residual Learning for Image Recognition'. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. DOI: <https://doi.org/10.1109/CVPR.2016.90>
- Heijnen, S, Lestrandoko, NH, Betjes, S, Hekster, O, Spreuwers, L and De Wit, F.** 2025. 'Beyond Hairstyle: Roman Imperial Portraits and Facial Recognition Software'. *International Journal for Digital Art History*, 10: 4.13–4.25. DOI: <https://doi.org/10.11588/dahj.2025.10.103033>
- Hertel, D.** 2013. *Die Bildnisse des Tiberius*. Das römische Herrscherbild I 3. Wiesbaden: Reichert.
- Kingma, DP and Ba, J.** 2017. *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs.LG]. <https://arxiv.org/abs/1412.6980>.
- Kunze, C.** 2015. 'Formal Approaches'. In: Marconi, C (Ed.) *The Oxford Handbook of Greek and Roman Art and Architecture*.

- Oxford: Oxford University Press, pp. 541–556. DOI: <https://doi.org/10.1093/oxfordhb/9780199783304.013.024>
- Kutrzyński, M and Król, D.** 2024. 'Deep Learning-Based Human Pose Estimation Towards Artworks Classification'. *Journal of Information and Telecommunication*, 8(4): 470–489. DOI: <https://doi.org/10.1080/24751839.2024.2331866>
- Langner, M.** 2021. 'Das Gesicht hinter dem Marmor. Computergestützte Maßvergleiche an römischen Kaiserporträts'. In: Lang, J and Marcks-Jacobs, C (Eds.) *Arbeit am Bildnis. Porträts als Zugang zu antiken Gesellschaften, Festschrift Dietrich Boschung*. Regensburg: Schnell & Steiner, pp. 377–384.
- Lestriandoko, NH, de Wit, F, Betjes, S, Heijnen, S, Hekster, O and Spreeuwiers, L.** 2024. 'The Role of Facial Hair on Roman Emperors' Face Recognition'. In: *12th International Workshop on Biometrics and Forensics (IWBF)*, pp. 1–7. DOI: <https://doi.org/10.1109/IWBF62628.2024.10593824>
- Lettinga, PE.** Nov. 2022. 'Recognizing Roman Emperors by Hair Using Several Machine Learning Methods'. thesis. University of Twente.
- Lin, T-Y, Dollár, P, Girshick, R, He, K, Hariharan, B and Belongie, S.** 2017. 'Feature Pyramid Networks for Object Detection'. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944. DOI: <https://doi.org/10.1109/cvpr.2017.106>
- Long, J, Shelhamer, E and Darrell, T.** 2015. 'Fully Convolutional Networks for Semantic Segmentation'. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440. DOI: <https://doi.org/10.1109/CVPR.2015.7298965>
- Lu, M, Zhang, Y, Zheng, B, Masuda, T, Ono, S, Oishi, T, Sengoku-Haga, K and Ikeuchi, K.** 2013. 'Portrait Sculptures of Augustus: Categorization via Local Shape Comparison'. In: *Proceedings of the 2013 Digital Heritage International Congress (DigitalHeritage)*. Piscataway: Institute of Electrical and Electronics Engineers, pp. 661–664. DOI: <https://doi.org/10.1109/DigitalHeritage.2013.6743812>
- Patay-Horváth, A.** 2014. 'Master-Hand Attributions of Classical Greek Sculptors by 3D-Analysis at Olympia – Some Preliminary Remarks'. In: Fellner, DW, et al. (Eds.) *GCH 2014. Eurographics Workshop on Graphics and Cultural Heritage*, Darmstadt, Germany, October 6–8, 2014. Goslar: Eurographics Association, pp. 41–42. DOI: <https://doi.org/10.2312/gch.20141327>
- Patay-Horváth, A and Christiansen, L.** Dec. 14, 2017. 'From Reconstruction to Analysis. Re-Use and Re-Purposing of 3D Scan Datasets Obtained from Ancient Greek Marble Sculpture'. *Studies in Digital Heritage*, 1(2): 491–500. DOI: <https://doi.org/10.14434/sdh.v1i2.23236>
- Ramesh, DS, Heijnen, S, Hekster, O, Spreeuwiers, L and De Wit, F.** 2022. 'Facial recognition as a Tool to Identify Roman Emperors: Towards a New Methodology'. *Humanities and Social Sciences Communications*, 9(78). DOI: <https://doi.org/10.1057/s41599-022-01090-y>
- Ronneberger, O, Fischer, P and Brox, T.** 2015. 'U-Net: Convolutional Networks for Biomedical Image Segmentation'. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer, pp. 234–241. DOI: https://doi.org/10.1007/978-3-319-24574-4_28
- Santos, J, Nunes, DAP, Padnevyh, R, Quaresma, JC, Lopes, M, Gil, J, Bernardes, JP and Casimiro, TM.** 2024. 'Automatic Ceramic Identification Using Machine Learning. Lusitanian Amphorae and Faience. Two Portuguese Case Studies'. *STAR: Science & Technology of Archaeological Research*, 10(1): e2343214. DOI: <https://doi.org/10.1080/20548923.2024.2343214>
- Selvaraju, RR, Cogswell, M, Das, A, Vedantam, R, Parikh, D and Batra, D.** 2017. 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization'. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626. DOI: <https://doi.org/10.1109/ICCV.2017.74>
- Sheng, Y.** 2024. 'Facial Recognition and Classification of Terracotta Warriors in the Mausoleum of the First Emperor Using Deep Learning'. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10: 205–212. DOI: <https://doi.org/10.5194/isprs-annals-X-1-2024-205-2024>
- Sokolova, M and Lapalme, G.** 2009. 'A Systematic Analysis of Performance Measures for Classification Tasks'. *Information Processing & Management*, 45(4): 427–437. DOI: <https://doi.org/10.1016/j.ipm.2009.03.002>
- Thakur, A, Biswas, SK, Majumdar, S and Thounaojam, DM.** 2025. 'A Comprehensive Review on Different CNN Architectures'. In: *3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. IEEE, pp. 1997–2004. DOI: <https://doi.org/10.1109/IDCIOT64235.2025.10914792>
- von den Hoff, R.** 2019. *Einführung in die Klassische Archäologie*. München: C.H. Beck. DOI: <https://doi.org/10.17104/9783406727290>
- Wisnangkol, J, Sanpechuda, T, Lim, S and Kovavisaruch, L-O.** 2024. 'Imbalanced Classification of Cultural Heritage Images Using Deep Neural Networks'. In: *First International Conference on Pioneering Developments in Computer Science & Digital Technologies (IC2SDT)*. IEEE, pp. 46–51. DOI: <https://doi.org/10.1109/IC2SDT62152.2024.10696516>
- Yang, W, Wei, Y, Wei, H, Chen, Y, Huang, G, Li, X, Li, R, Yao, N, Wang, X, Gu, X, others.** 2023. 'Survey on Explainable AI: From Approaches, Limitations and Applications Aspects'. *Human-Centric Intelligent Systems*, 3(3): 161–188. DOI: <https://doi.org/10.1007/s44230-023-00038-y>
- Zhang, Y, Lu, M, Zheng, B, Masuda, T, Ono, S, Oishi, T, Sengoku-Haga, K and Ikeuchi, K.** Sept. 2013. 'Classical Sculpture Analysis via Shape Comparison'. In: *International Conference on Culture and Computing*. Kyoto, Japan: IEEE, pp. 57–61. DOI: <https://doi.org/10.1109/CultureComputing.2013.18>

TO CITE THIS ARTICLE:

Bergmann, B, Matuszczyk, D, Fisseler, D, Schröder, T, Schröder-Griebel, N and Weichert, F. 2026. Investigation Into Deep Learning Methods for a Formal Analysis of Ancient Sculptures. *Journal of Computer Applications in Archaeology*, 9(1): 400–417. DOI: <https://doi.org/10.5334/jcaa.238>

Submitted: 04 August 2025 **Accepted:** 28 May 2026 **Published:** 16 October 2026

COPYRIGHT:

© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <https://creativecommons.org/licenses/by/4.0/>.

Journal of Computer Applications in Archaeology is a peer-reviewed open access journal published by Ubiquity Press.