Learning Diachronic Representations of Ancient Greek Letterforms

John Pavlopoulos[1,2⋆], **Spyros Barbakos**[2,3], **Lavinia Ferretti**[4,5], **Dionysis Voulgarakis**[2], **Asimina Paparrigopoulou**[6], **Maria Konstantinidou**[6] **Giuseppe De Gregorio**[5,7], **Isabelle Marthot-Santaniello**[5], **Paraskevi Platanou**[1,2], **Holger Essler**[8]

[1] Athens University of Economics and Business, Greece ipavlopoulos@aueb.gr
[2] Archimedes, Athena Research Center, Greece
[3] Department of Computer and Systems Sciences, Stockholm University, Sweden
[4] Università degli Studi di Torino, Italy
[5] University of Basel, Switzerland
[6] Democritus University of Thrace, Greece
[7] Computer Vision Center (CVC) - Barcelona, Spain
[8] Julius-Maximilians-Universität Würzburg, Germany

Abstract. Learning representations that remain robust across centuries of variation in handwriting is a key challenge in diachronic representation learning. Taking one of the longest continuously used writing systems, ancient Greek, as a case study, we introduce three datasets for diachronic representation learning: **Hell-Char**, a curated training set spanning the 3rd–1st centuries BCE, and two evaluation sets, **PaLit-Char** (2nd–5th c. CE) and **Med-Char** (9th-14th c. CE). To address the challenges of symbolic variation, scarce data, and systematic degradation, we propose: a *similarity-weighted supervised contrastive loss* that biases embeddings using dynamically estimated inter-class similarities, and a *lacuna-driven augmentation* scheme that simulates realistic manuscript corruptions. Trained with these strategies, both a lightweight CNN and a pretrained ResNet achieve strong recognition performance and produce embeddings that more coherently separate character classes than PCA or generic pretrained models. These embeddings enable clustering, identification of stylistic subgroups, and construction of prototype images that visualize diachronic evolution and transitional letterforms. Our results demonstrate that respecting intrinsic inter-letter relationships and augmenting with domain-informed corruptions yield robust, interpretable representations, offering a transferable paradigm for representation learning under scarce, temporally evolving, and noisy conditions. Code and data available at: https://github.com/ipavlopoulos/diachronic-greek-letterforms.

Keywords: Diachronic Representation Learning · Historical Document Analysis · Handwritten Character Recognition · Clustering.

⋆ JP undertook the experiments and wrote the first draft; MK, GDG, IMS oversaw experiments and co-authored. LF co-authored (esp. §4 and §6) and provided Hell-Char and expertise. PP and HE co-authored and provided expertise. AP provided MedChar. SB and DV assisted with experiments and conceptualisation.

1 Introduction

Paleographic analysis of historical scripts requires robust automated character representation, a challenge that remains particularly acute for scripts such as ancient Greek. Greek handwriting spans over two and a half millennia, encompassing formal literary hands and highly cursive scripts, with substantial variation in stroke shape, scale, slant, and contextual noise [4, 6, 12, 1]. Material degradation and heterogeneous digitization practices further compound these challenges, introducing ambiguities that complicate segmentation, feature extraction, and character recognition and classification, especially under limited and imbalanced datasets. Although sometimes considered a low-level task, automated character representation has a significant impact for broader paleographic analysis, supporting text-image alignment, semi-automatic transcription, and tasks such as script typology, dating, and scribal attribution.

Existing document analysis and recognition methods typically assume stable character morphology and sufficient training data. These assumptions break down in historical scripts, where letterforms evolve gradually across centuries and exhibit systematic structural drift. Hence, standard transfer and contrastive learning approaches do not explicitly model structured inter-class similarity, treating visually related but distinct letterforms as equally dissimilar negatives.

This study addresses this challenge by learning robust character representations, with diachronic generalization serving as the central test case. We focus on the evolution of ancient Greek letters through a representation learning framework that incorporates paleographic knowledge. We propose two domain-driven methodological innovations: a *lacuna-driven augmentation* (LF) that simulates realistic manuscript degradations, and a *similarity-weighted* supervised contrastive loss (DSCL) that reweights negative pairs according to dynamically estimated inter-class similarities. LF targets the non-rectangular loss patterns produced by manuscript damage, while DSCL targets morphologically confusable letter classes that should not be treated as uniformly distant negatives. We evaluate their impact on both recognition performance and representation structure. Confusion matrices reveal systematically easy and difficult letters, while clustering analyses on the learned embeddings uncover stylistic subgroups and visually coherent forms. Prototype visualizations per letter–century further enable quantitative and interpretable analysis of diachronic evolution. Compared to raw pixels, PCA, or generic pre-trained features, our embeddings yield more coherent and discriminative representations of historical Greek handwriting.

We summarize our contributions as the following four key points:
1. We propose a representation learning objective that, unlike standard SCL or hard-negative mining, explicitly models inter-class similarity structure, preventing repulsion between inherently similar graphemes.
2. We introduce a domain-informed augmentation scheme that simulates manuscript degradations (lacunae) more faithfully, increasing robustness to missing or corrupted strokes.
3. We introduce historical Greek handwriting datasets spanning from the 3rd century BCE to the 14th century CE: Hell-Char (3rd–1st BCE), de-

rived from Hell-Date for character-level training and benchmarking; and two newly compiled evaluation datasets, PaLit-Char (2nd–5th CE) and Med-Char (9th–14th CE) for testing generalization across temporal shifts.
4. We conduct computational paleographic analyses using CNN-derived embeddings. We perform clustering, silhouette-based subgroup detection, and prototype visualization per letter–century, providing interpretable insights into diachronic variation and scribal conventions.[1]

2 Related Work

We are not aware of any other study in the literature that analyses the diachronic evolution of Greek handwritten letters from Antiquity to pre-modern times using machine learning. However, we acknowledge the existence of related fields, such as optical character recognition (OCR), and of other investigations on Greek papyri at the character level, which we discuss next.

OCR. Early OCR approaches relied on manual feature extraction methods, such as zoning, projection histograms, and contour profiling, to distinguish between characters. A comprehensive survey emphasized the importance of these handcrafted features in OCR [22], while in [10] a grapheme-based feature extraction system was introduced that modelled diachronic variations while incorporating textual features. The advent of deep learning has further transformed the field. In [14], the authors demonstrated the effectiveness of convolutional neural networks (CNNs) in classifying handwritten digits, laying the groundwork for modern neural approaches in character recognition. Autoencoders [11] and contrastive learning [5] have gained traction in unsupervised learning, enabling models to learn meaningful representations of handwriting directly from data, without the need for manual feature engineering. Building on these advances, several of the latter works have examined deep learning for feature analysis of some aspects of ancient Greek handwritings. In [16], the authors addressed the issue of clustering historical handwriting by similarity with no metadata explicitly indicating date or style. Their method strongly focuses on character, employing a SimSiam neural network to quantify similarity between images of single Greek letters (Alpha, Epsilon, and Mu) from different manuscripts. Their observations of stylistic similarity are useful to paleographers as they situate manuscripts within an integrated network and disclose subtle micro-phenomena of similarity.

CNNs. CNNs have been applied to OCR-extracted text [15], combining visual and textual features to improve dating accuracy. However, their approach assumes the availability of high-quality (historical yet printed) data conducive to accurate OCR results, an assumption that often fails in the context of historical documents such as the Greek papyri addressed in our study. To tackle

[1] Code and data are available at: https://github.com/ipavlopoulos/diachronic-greek-letterforms.

such challenges, an ImageNet-pretrained CNN was fine-tuned on a corpus of medieval documents, demonstrating improved performance on degraded or irregular scripts [23]. More chronology-specific, in [24], a deep learning pipeline was designed for the automated dating of images of ancient Greek papyrus fragments. A multi-stage pipeline integrated handwritten text recognition (HTR) for character detection and classification, followed by distinct character-level and fragment-level dating prediction models. Their aggregated sum of fragment-level models reaches up to 79% accuracy in predicting two-century-broad date ranges on fragments with large numbers of characters. More recently, a transformer-based pipeline was proposed [2] that integrates classical preprocessing techniques with a fine-tuned Vision Transformer and majority voting for document dating. This study pioneers the integration of Vision Transformers in the context of historical manuscript dating, a domain where CNNs were dominating.

SimCLR. A simple yet powerful contrastive framework for representation learning was introduced in [5]. Each image is augmented twice, and the network is trained to maximize agreement between positive pairs while treating all other samples in the batch as negatives. While effective as a simple self-supervised technique at scale, SimCLR assumes that all non-matching samples are equally dissimilar. In fine-grained recognition tasks such as character classification, this uniform treatment forces visually similar but distinct classes apart (e.g., A vs. Λ), discarding useful structural information.

SCL. In [13], the authors extended SimCLR to the labelled setting by grouping all samples of the same class as positives. This produces tighter class-specific clusters. Importantly, they also showed that combining supervised contrastive embeddings with a linear classifier trained under cross-entropy further improves classification accuracy compared to cross-entropy alone. However, SCL still treats all negatives uniformly, regardless of their visual similarity to the anchor. As a result, classes with inherent affinities (e.g., letters with similar shapes) are repelled too strongly, yielding embeddings that fail to reflect natural inter-class relationships. In addition to instance discrimination, weakly SCL [26] introduced a supervised contrastive component based on weak labels derived from K-nearest neighbor graphs. Instead of treating all other samples as negatives, this approach dynamically identifies semantically similar neighbors and reweights them as positives, alleviating the class collision problem. SCL treats negatives uniformly and makes classes with inherent affinities (e.g., letters with similar shapes) to be strongly repelled. This leads to embeddings that fail to reflect natural inter-class relationships. Our study addresses this gap.

3 Methodology

We analyse handwritten Greek letters across centuries using CNN-based embeddings enhanced with advanced fragmentation and SCL strategies.

3.1 CNN Backbone and Lacunae Fragmentation

A 2D CNN (fCNN) was suggested in [18] for dating images of papyrus lines, which comprised a fragmentation-based augmentation strategy. We follow a similar fragmentation-based strategy, yet our CNN differs in two ways. First, it is adjusted to operate on letters instead of text lines. Second, the fragmentation augmentation is improved so that synthetic lacunae follow their natural (curvy) shape, i.e., circular or elliptic, not square (§5.1). The trained model produces high-dimensional embeddings $\mathbf{e} \in \mathbf{R}^D$ representing the visual structure of each letter. The base CNN architecture consists of convolutional layers to extract local stroke and shape patterns; ReLU activations for non-linearity; pooling layers to reduce spatial dimensions while preserving salient features; fully connected layers to map feature maps into the final embedding vector. These embeddings abstract style variations while preserving essential letterform characteristics. We also experiment with ResNet18 pre-trained CNN [9], the ConvNext-V2 self-supervised and globally-normalized CNN [25], and the ViT-S16 Vision Transformer [3].

3.2 Augmentation

Each character image is converted to grayscale, normalized, and resized to 64×64 pixels. To account for variability in handwriting and material degradation, we applied rotation (up to 10°), translation, resizing, color jittering, and lacunae-inspired masking (LF), which simulates missing ink or manuscript damage, improving the model's robustness to partial character visibility (Fig. 1, rightmost).

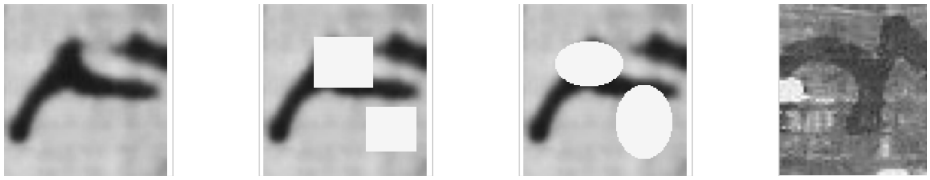

Fig. 1: Alpha from Hell-Char (leftmost) with rectangular (RE, 2nd) and lacuna-inspired (LF, 3rd) masking. Alpha with naturally fragmented surface (4th).

3.3 Similarity-Weighted Supervised Contrastive Loss

In addition to the standard cross-entropy loss (i.e., the supervised letter-classification objective applied to the backbone's classification head), we train the backbone models using a SCL loss, which encourages embeddings of the same letter to cluster together while pushing apart visually dissimilar letters. Visual similarities between letters, dynamically learned,[2] are used to weight negative pairs,

[2] Visual similarities could also be defined manually, but our experiments (using a prior similarity matrix based on modern letter shapes) did not lead to improvements.

enabling the model to respect intrinsic inter-letter relationships. This contrastive loss is not computed on the classification logits, but it is applied to the intermediate feature embeddings produced by the backbone before the classification layer. Thus, the model jointly optimizes cross-entropy on the classification head and contrastive loss on the shared backbone representations. For each anchor embedding $\mathbf{e}_i$, the loss is defined as:

$$\mathcal{L}_i = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{e}_i \cdot \mathbf{e}_p / \tau)}{\sum_{a \neq i} w_{ia} \exp(\mathbf{e}_i \cdot \mathbf{e}_a / \tau)}$$

where $P(i)$ is the set of positive samples (same class as i) and τ is the softmax temperature; $w_{ia} = 1 + \lambda \frac{S_{y_i, y_a}}{\bar{S}}$ is the weight for negative pair (i, a); S_{y_i, y_a} is the similarity between classes y_i and y_a, dynamically computed from embeddings; $\bar{S}$ is the mean off-diagonal similarity; and λ controls the influence of similarity weighting. This loss ensures that embeddings of the same letter cluster tightly, while visually similar letters exert weaker repulsion.

3.4 Prototype Selection (Medoid)

For each group (letter, century), we select a representative *medoid* embedding to serve as a prototype (T), defined as: $T = \arg\min_i \sum_{j=1}^{N} \big(1 - \cos(\mathbf{e}_c, \mathbf{e}_j)\big)$, where N is the number of embeddings in the group and e_c is the centroid. The medoid provides a highly representative image that is robust to outliers.

4 Dataset Development

4.1 Source

The Hell-Date dataset [8] comprises 194 images from 157 papyri, all written in Greek and dated between the years 310 BCE and 3 BCE. The material is particularly relevant for digital paleography and papyrological analysis due to its historical span, script diversity, and accompanying metadata. Each document in the dataset is associated with rich contextual metadata, including the date of composition, the geographical provenance, and the textual type. Of the 194 available images, 171 are annotated at the character level, forming the primary subset of character images used in this work. We used this character-level subset but filtered and restructured it for our purposes; we refer to the restructured subset as Hell-Char (see below). This is the first study to utilize the character annotations included in Hell-Date, which are further presented below.

The character annotations in Hell-Date. Twenty-nine character classes are present in the annotations of Hell-Date. In addition to the 24 standard letters of the Greek alphabet, the dataset comprises 3 archaic numeral letters (*Stigma*, *Qoppa*, and *Sampi*). It also uses a general 'symbol' category for all characters that are not alphabetic letters. Last, an 'unknown' class was added for uncertain or ambiguous signs, but it remained empty. Each character instance is also assigned

a base-type (BT) tag, ranging from BT1 to BT5, which indicates its degree of preservation. BT1 is the best-preserved, but all can be useful in the future to analyze the correlation between physical degradation and recognition performance.

4.2 The Hell-Char subset

As said above, Hell-Char is a subset of Hell-Date annotations. The classes for archaic numerals, symbols, and unknown letters were merged into a single general non-alphabetic category labelled "Unknown", which was ignored in later stages of our work. We also limited our analysis to letters tagged BT1. To reduce the imbalance in character frequency and to ensure a more uniform distribution of samples across classes and documents, at most five instances per character class per papyrus were randomly selected. The resulting subset comprises 13,046 character images from 157 distinct papyri. Figure 2 shows the letter frequency in Hell-Char.

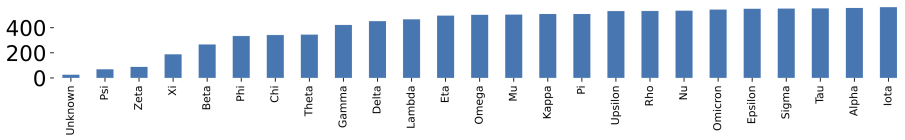

Fig. 2: Letter frequency in our Hell-Char subset.

Intrinsic challenges. The dataset presents several intrinsic challenges. The character class distribution is still unbalanced (Figure 2) due to the large variation in letter frequency in Greek (the highest frequency is over 10 percent for Alpha, while below 0.5 percent for the three rarest letters Psi, Zeta and Xi). Besides, some glyphs are ambiguous even to experts when considered without any linguistic or chronological context. These characteristics make Hell-Char a valuable, non-trivial benchmark for evaluating character recognition in ancient scripts.

5 Empirical Analysis

5.1 Experimental Settings

The Similarity Matrix. We re-estimate the class-similarity matrix periodically (every 3 train epochs). At each update, we pass the entire training set through the current model, compute class prototypes from the normalized embeddings, and derive the cosine similarity between prototypes. Cosine similarities are clamped to $[0, 1]$ and diagonal entries are set to zero. This yields a dynamic measure of inter-class similarity that evolves with the representation space; an exponential moving average can be applied to stabilize updates. The updated matrix is then used by our DSCL, which down-weights negatives from highly similar classes and up-weights negatives from dissimilar ones.

Lacuna-driven Synthetic Fragmentation. We attempt to simulate manuscript deterioration more realistically than standard erasure augmentations by inserting irregular regions that approximate actual lacunae observed in historical documents. For each image, we sample 1–4 lacunae and each covers 2–15% of the area to match the typical size distribution of physical papyrus damage. Lacuna geometry is obtained by drawing anisotropic ellipses whose contours are further distorted via random morphological operations (erosion or dilation), producing organic, non-rectangular shapes characteristic of flaking, humidity damage, surface wear, or other deterioration (e.g., worm holes are frequent in papyri). These lacunae are placed at random positions and the masked pixels are replaced with background values, reflecting the absence of ink/papyrus rather than additive noise. This augmentation increases robustness to fragmentary handwriting and introduces realistic variability at negligible computational cost.

Data Split. We keep 20% of the data for testing, following a letter-based stratified split. Although we acknowledge that this strategy allows a scribe-based leakage, the selected approach better fits the scope of this work (see §7).

5.2 Letter Recognition

Table 1 shows the performance of backbones when we add our LF and DSCL. The

Table 1: Classification performance on Hell-Char (sorted) of fCNN [18] and ResNet18 [10], pre-trained (PT) and/or fine-tuned (FT), when we add: our SCL with dynamically-learned weights (DSCL), and fragmentation-based augmentation: i.e., none (-), random erasure (RE), or our lacuna-driven (LF).

Model	Fragmentation	Contrastive Loss	**Accuracy**	**F1**
fCNN	-	-	0.742	0.74
fCNN	RE	-	0.768	0.75
fCNN	LF	-	0.782	0.77
ResNet18-FT	-	-	0.788	0.74
ResNet18-PT+FT	-	-	0.801	0.79
ResNet18-PT+FT	-	SCL	0.818	0.81
fCNN (ours)	LF	DSCL	0.803	0.80
ResNet18-PT+FT (ours)	LF	DSCL	0.831	**0.83**

vanilla fCNN [18] without any fragmentation, achieves an Accuracy of 74%. F1 is exactly the same, indicating the balanced performance across letters despite the class imbalance (Figure 2). Random erasure (RE) [18], shown in Fig. 1, leads to improvements in both metrics. Our LF, however, outperforms both. The architecture of ResNet18 [10] performs worse in F1 (on par in Accuracy) when

trained from scratch, and improves when pre-trained. Standard SCL improves further ResNet18-PT+FT, which is outperformed only by our LF+DSCL version (0.83). When we enhance fCNN with our LF and DSCL, it outperforms the pre-trained ResNet18. Preliminary experiments show that LF and DSCL also improve the classification performance of ViT-16S and ConvNeXt-V2, but an analytical investigation of the gains across backbones is left for future work.

5.3 Letter Image Clustering

Our previous experiments showed that CNN-based embeddings can be used to represent letters. To assess the quality of the resulting embeddings, we compared them against baseline features, then feeding algorithms that should cluster images of the same letter into subcategories. We also engineered features, based on Otsu's method [17], a widely used adaptive thresholding technique, and principal component analysis (PCA) [19], keeping as many dimensions as add up to 90% of the original information (i.e., 500). Characters have consistent alignment and size, hence pixel-based variance captured by PCA can correspond to meaningful features of the characters (e.g., strokes and overall shape). Although it destroys 2D structure (edges, texture) and does not focus on separability, PCA applied to the raw input is a simple preprocessing baseline that is complementary to CNN features that preserve local structure. The empirical results, shown in Table 2,

Table 2: Clustering performance on Hell-Char using different embeddings and clustering algorithms, sorted by performance across algorithms.

	k-means		Spectral		AH	
Embedding	NMI	ARI	NMI	ARI	NMI	ARI
ResNet18+LF+DSCL	**0.768**	**0.690**	**0.765**	**0.683**	**0.760**	**0.674**
fCNN+LF+DSCL	0.428	0.189	0.631	0.442	0.544	0.292
ResNet18+PT+FT	0.480	0.257	0.487	0.225	0.464	0.197
Otsu+PCA	0.318	0.152	0.382	0.176	0.356	0.168
ResNet18+PT	0.067	0.010	0.094	0.015	0.073	0.008

underscore the importance of task-specific embeddings in historical handwriting clustering. ResNet18, fine-tuned and enhanced with LF and DSCL, consistently outperforms both Otsu+PCA and a pre-trained ResNet18, achieving markedly higher agreement with paleographic labels across all metrics. Otsu+PCA, though superior to raw pretrained features, lags far behind, while ResNet18 fails entirely, with near-random partitions. The stark contrast highlights two key findings: (i) general-purpose CNN features trained on ImageNet-style photographs are far from isolated historical character images and do not transfer directly to paleographic tasks, and (ii) the manifold structure of handwritten letter embeddings is not captured adequately by centroid-based partitioning. Together, these results

validate the need for domain-tailored architectures and manifold-aware clustering to recover meaningful structure in diachronic handwriting data.

5.4 Pattern Recognition: Revealing Letter Forms

Using Spectral Clustering on the embeddings of our best performing CNN (i.e., ResNet18+lf+dscl; see Table 1), we applied the Silhouette method [21] to detect the optimal number of clusters per letter. For each letter, we varied the number of clusters and retained the configuration with the highest Silhouette score [20]. For Alpha, the optimal number exceeded the two clusters indicating multiple distinct forms,[3] shown in Figure 3. Clustering letter forms into subtypes

Fig. 3: Representative forms of letter Alpha using cluster medoids (§3.4).

is a hard and unsolved task in paleography. The network's results may, in some cases, point to useful characteristics. In the case of Alpha, the three images seem to represent different characteristics: the first alpha is filled-in (no empty space in its centre), circular and ligatured to the left; the second one is circular but not ligatured; the third one is angular. This partition is coherent from a paleographical point of view.

6 Out of Temporal Distribution Application

Our backbone CNNs are trained on letter images from papyri of the last three centuries BCE. In this period, the epigraphic letter forms (close to our modern capital letters) start to be modified with increasing cursivity, driven by the practical demands of faster writing. This increase in cursivity continues over the following centuries and constantly deforms letter shapes; however, epigraphic letter forms are maintained, especially for calligraphic writing styles called capital (or uncial) bookhands. In parallel, a calligraphic stylization of cursive forms gradually develops until it reached the so-called state of "minuscule script" in the 9th c. While many minuscule letterforms remain visually close to their capital ancestors, others diverge significantly (e.g. Beta, Mu, Gamma, and Delta). During the following centuries, uncial and minuscule calligraphic forms continued to coexist, sometimes within the same manuscript and even within a single word.

[3] Silhouette scores cannot be computed for a single cluster; hence the minimum number considered was two.

6.1 Evaluation Dataset Development

PaLit-Char: Majuscule Literary Papyri. To evaluate how well the model generalizes to letterforms close in time to the training data, we constructed the **PaLit-Char** test set. It is a fully balanced dataset containing 384 images (4 specimens × 24 letters × 4 centuries) spanning the 2nd–5th CE. Images were drawn from securely dated literary papyri in the PaLit dataset [18]; for the 5th century, where securely dated material is scarce, 48 images were taken from an additional, paleographically dated manuscript. While Hell-Char covers cursive handwriting from the last three centuries BCE, PaLit-Char extends into the early centuries CE and covers calligraphic writing, offering both chronological continuity and stylistic diversity. This allows us to test whether features learned on late Hellenistic cursive letters transfer to Roman-period bookhands that retain strong ties to their predecessors but already display variation.

Med-Char: Medieval Minuscule Manuscripts. With the historical evolution described above in mind—and having first tested the recognition performance of our network on the chronologically close PaLit-Char—we proceed to test its ability to recognize letterforms from medieval minuscule manuscripts. This evaluates both the generalizability of learned features across paleographic periods and the limits of shape-based classification given the diachronic script variation. To assess this hypothesis, we compiled a dataset of 574 letter images from manuscripts dated between 835 and 1378 CE, a much later period. We used 24 images per letter, opting for balance across the centuries in that period,[4] and using the best performing ResNet18, enhanced with our LF and DSCL, to classify each image. We call this evaluation dataset **Med-Char**. Contrary to our training set and the PaLit-Char test set, which contain capital or cursive letters (upper case), Med-Char is a Byzantine minuscule letter (lower case) dataset. This choice is deliberate: minuscule script is historically derived from majuscule but exhibits substantial graphic divergence, with some letters retaining visual continuity and others undergoing radical transformation. Testing on Med-Char, hence, allows us to probe the limits of the learned representations under extreme diachronic and stylistic shift, providing a benchmark for cross-period generalization.

6.2 Experimental Analysis

Closer in Time. On PaLit-Char (Table 3), ResNet18+LF+DSCL achieves an Accuracy and F1 of 0.84, very close to the results of Hell-Char. Although F1 decreased for specific letters (e.g., Phi, Pi, Psi), it improved for others (e.g., Alpha and Zeta). The calligraphic nature (regular, standardized, legible) of PaLit characters can explain this increase.

[4] We include 24 random instances of each letter per century from multiple manuscripts. The rarest letter, Psi, has only 22 occurrences.

Table 3: Diachronic generalization of ResNet18+lf+dscl across datasets.

Test Set	Period	Accuracy	F1-score
Hell-Char	3rd–1st c. BCE	0.83	0.82
PaLit-Char	2nd–5th c. CE	0.84	0.84
Med-Char	9th–14th c. CE	0.45	0.42

Far Away in Time. The same model reaches 0.45 Accuracy on Med-Char with highly uneven per-class performance. Chi, Lambda and Iota are recognised best (F1 0.87, 0.80, 0.77), as their capital, cursive and minuscule shapes remain similar. In contrast, Alpha, Gamma and Upsilon collapse to zero F1, reflecting systematic confusion with visually similar shapes and high diachronic variability that undermines generalization. Gamma, for instance, undergoes a strong visual evolution: in Hell-Char, its shape is still close to epigraphic Gamma Γ, whereas in Med-Char, it is close to minuscule Gamma γ, a shape that resembles epigraphic Upsilon of Υ in Hell-Char. The remaining letters trade Precision against Recall: Omicron, Tau, and Kappa are over-predicted (Recall 1.00, 0.96, 0.88, but Precision 0.23, 0.48, 0.41), whereas Psi is under-predicted (Precision 0.90, Recall 0.41), and Delta is almost never predicted, but in the rare cases it is, it is correct. As can be seen in Figure 4, misclassification patterns are temporally structured.

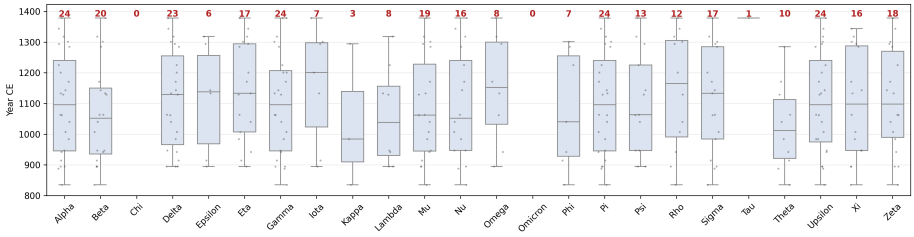

Fig. 4: Boxplot of years per letter for misclassified out-of-distribution images of Med-Char. Red numerals indicate the number of mistakes.

Best-recognised Chi and over-predicted Omicron have no (Recall) errors, and Tau only a single one (near 1378 CE). For most letters the errors concentrate in the 11th–12th centuries (median around 1050–1130 CE). The earliest confusions involve Kappa (around 985 CE) and the latest Iota (around 1200 CE), implying that historical morphological shifts exert non-uniform effects on recognition difficulty. Noteworthy is the fact that fine-tuning on PaLit-Char and inferring on Med-Char brings only modest gain, from 0.45 to 0.48 Accuracy. This transfer remains difficult because Roman and late-antique majuscule bookhands differ structurally from medieval minuscule, and the intermediate cursive stages that connect them are absent from the training data.

Letter-Century Clusters. Figure 5 illustrates a two-dimensional t-SNE projection of the ResNet18+lf+dscl embeddings, where each point corresponds to an image patch representing a handwritten Greek Med-Char character. To reduce clutter and improve interpretability, instead of showing all individual samples, one prototype image per letter–century pair is overlaid: the prototype is chosen as the sample closest to the centroid of its group in the t-SNE space, thus representing the most "typical" example of that cluster. The resulting map highlights how temporal and graphemic factors shape the embedding space. Within the clusters related to one character, overlapping or diffuse areas indicate stylistic continuities or transitional forms between centuries, whereas sharp separations reveal periods of stronger diachronic variation. This approach provides an interpretable way of assessing the alignment between automated embeddings and paleographic expectations, enabling both qualitative validation of the clustering behavior and the identification of anomalies or particularly distinctive exemplars.

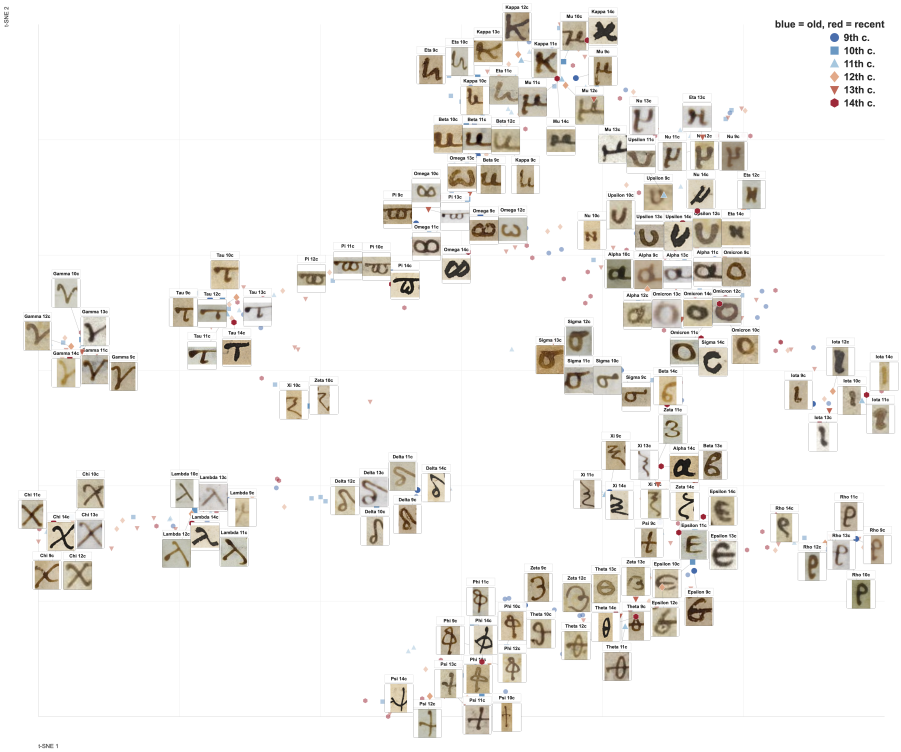

Fig. 5: t-SNE plot of ResNet18+lf+dscl embeddings on Med-Char. Points are coloured by century (blue for older, red for more recent). One prototype image per letter-century group is shown, selected as the sample closest to the group centroid, to visualize graphemic clustering and local diachronic variation.

Distinctively Isolated Letters. Several letters that obtained high F1 scores (e.g., Iota ι, Lambda λ, Chi χ) form distinct peripheral clusters, consistent with their relatively stable shapes across the training material. Gamma also forms an isolated cluster, but its form varies greatly from Hellenistic to Medieval times. This strong diachronic shift can explain why the model fails to recover Gamma in Med-Char: true medieval Gamma examples no longer match the Gamma forms learned from Hell-Char, leading to zero Precision and Recall for the class.
Visible Clusters: The cases of Zeta and Beta. Multiple forms of Zeta are visible, including examples whose shape resembles a 3 and others closer to modern-day ζ, often near letters with related stroke structure such as Theta or Xi. This distinction points to one reason for confusion in recognizing Zeta: its "3-looking" shape is absent from the training data and therefore, cannot be identified. The same is true for Beta: its B-shaped and minuscule, u-shaped forms occupy different neighbourhoods of the embedding space, with the latter appearing near other minuscule u-shaped letters such as Kappa, Mu and Eta. This u-shaped Beta, absent from Hell-Char, explains its very low Recall.
Typical Medieval Forms A broad region of the graph groups typical Medieval letter forms based on successions of 'o' and 'u' shapes. These shapes are quite different from Hell-Char letters shapes, and indeed some of the letters represented here (Omega, Beta, Kappa, Mu, Nu, Upsilon) do not belong to the top-performing ones. Kappa and Nu achieved an F1 of 0.56 and 0.47 respectively, which can be explained by their mixing Medieval, minuscule, u-shaped forms with older, capital forms already attested in Hell-Char. These older K and N forms appear near the edges of this broader minuscule-form neighborhood.

7 Discussion

Novel Methodological Contributions. Our core innovations lie in the LF augmentation scheme and DSCL similarity-weighted loss. The empirical results clearly demonstrate that these methods significantly outperform standard CNN baselines and generic pre-trained models, which often fail to transfer to the variability of paleographic data. Erasing input as augmentation in image classification is not new [27], and it has been shown to be particularly useful for papyri, which are often fragmented. **Our presented LF augmentation** is closer in nature to the real fragments compared to RE (see Figure 1), while **our proposed DSCL** yields embeddings that reflect natural inter-class relationships. By contrast, standard SCL treats negatives uniformly, making classes with inherent affinities (e.g., letters with similar shapes, such as Psi-Phi) to be strongly repelled.

Scribe Leakage and Diachronic Generalization. We assess the robustness of our representations by testing the model on chronologically distinct datasets (Table 3). The stable performance on PaLit-Char (0.84 F1) confirms that the model captures structural letterforms rather than individual handwriting styles, effectively mitigating the risk of scribe leakage. The performance drop observed on Med-Char (0.42 F1) further reinforces this defense: it reveals that the model is

sensitive to the radical morphological shift from Hellenistic cursive to Byzantine minuscule (e.g., the evolution of Γ into γ). Our error analysis shows that the model, lacking intermediary Roman-period cursive data, maps unknown minuscule shapes to the closest visual proxies, such as confusing Med-Char Gamma with Upsilon. This confirms that our framework successfully learns transferable paleographic features that follow historical evolutionary paths rather than memorizing specific training samples. While the low sample density (max 120 characters per papyrus) remains a challenge for scribal identification [7], it forces the model to focus on diachronic morphological stability.

8 Conclusions

We introduce three datasets of historical Greek handwriting (Hell-Char, PaLit-Char, Med-Char), using them to examine how modern representation learning captures symbolic variation across time. Beyond establishing Hell-Char as a benchmark for low-resource, domain-shifted visual recognition, we propose two methodological innovations: a similarity-weighted supervised contrastive loss, which explicitly models structured inter-class similarity, and a lacuna-driven augmentation scheme, which simulates realistic manuscript degradations. Empirically, we demonstrate that CNN-derived embeddings yield more discriminative and structured representations than PCA or generic pre-trained features, while clustering uncovers stylistic subgroups that mirror diachronic variation and coexisting scribal conventions. Prototype visualisations per letter–century further illustrate gradual graphical change, providing interpretable bridges between computational analysis and paleographic interpretation. More broadly, our findings highlight the limits of natural-image transfer learning for historical document analysis and show that incorporating domain-informed similarity and degradation modelling improves robustness under temporal domain shift.

Acknowledgements

This work has been partially supported by: project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program; and the Swiss National Science Foundation, SNSF Starting Grant TMSGI1-211682 "EGRAPSA: Retracing the evolutions of handwritings in Greco-Roman Egypt thanks to digital paleography."

This version of the contribution has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/978-3-032-36039-7_10. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/acceptedmanuscript-terms.

Bibliography

[1] Bianconi, D.: Greek Palaeography. In: Bausi, A., Borbone, P.G., Briquel-Chatonnet, F., Buzi, P., Gippert, J., Macé, C., Maniaci, M., Melissakis, Z., Parodi, L.E., Witakowski, W. (eds.) Comparative Oriental Manuscript Studies. An Introduction, pp. 297–305. Trendition, Hamburg (2015)

[2] Boudraa, M., Bennour, A., Al-Sarem, M., Ghabban, F., Bakhsh, O.A.: Contribution to historical manuscript dating: A hybrid approach employing hand-crafted features with vision transformers. Digital Signal Processing **149**, 104477 (2024). https://doi.org/10.1016/j.dsp.2024.104477

[3] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 4778–4788 (2021)

[4] Cavallo, G.: Greek and Latin Writing in the Papyri. In: Bagnall, R.S. (ed.) The Oxford Handbook of Papyrology, pp. 101–148. Oxford University Press, Oxford, New York (2009)

[5] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of ICML (2020). https://doi.org/10.48550/arXiv.2002.05709

[6] Crisci, E., Degni, P.: La scrittura greca dall'antichità all'epoca della stampa. Una introduzione. Carocci, Roma (2011)

[7] D'Alessandro, T., Cilia, N., De Stefano, C., Fontanella, F., Molinara, M., Scotto di Freca, A., Marthot-Santaniello, I.: A deep transfer learning approach for writer identification in greek papyri. ACM Journal on Computing and Cultural Heritage **18**(3), 1–19 (2025)

[8] Ferretti, L., Serbaeva Saraogi, O., De Gregorio, G., Marthot-Santaniello, I.: Hell-Date. EGRAPSA Hellenistic Dated Papyri Dataset (2025). https://doi.org/10.5281/zenodo.15083590, https://zenodo.org/records/15083590

[9] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)

[10] He, S., Schomaker, L.R., Samara, P., Burgers, J.: Mps data set with images of medieval charters for handwriting-style based dating of manuscripts. https://doi.org/10.5281/zenodo.1194357 (2016)

[11] Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006). https://doi.org/10.1126/science.1127647

[12] Irigoin, J.: L'alphabet grec et son geste des origines au IXe siècle après J.-C. In: Sirat, C., Irigoin, J., Poulle, E. (eds.) L'écriture: le cerveau, l'œil et la main, pp. 299–305. Brepols, Turnhout (1990)

[13] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: NeurIPS. pp. 18661–18673 (2020), proceedings.neurips.cc

[14] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791
[15] Li, Y., Genzel, D., Fujii, Y., Popat, A.C.: Publication date estimation for printed historical documents using convolutional neural networks. In: The 3rd International Workshop on Historical Document Imaging and Processing. pp. 99–106. ACM (2015)
[16] Marthot-Santaniello, I., Vu, M.T., Serbaeva, O., Beurton-Aimar, M.: Stylistic similarities in greek papyri based on letter shapes: A deep learning approach. In: Coustaty, M., Fornés, A. (eds.) ICDAR Workshops. pp. 307–323. Springer Nature Switzerland, Cham (2023)
[17] Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (1979). https://doi.org/10.1109/TSMC.1979.4310076
[18] Pavlopoulos, J., Konstantinidou, M., Perdiki, E., Marthot-Santaniello, I., Essler, H., Vardakas, G., Likas, A.: Explainable dating of greek papyri images. Machine Learning **113**(9), 6765–6786 (2024)
[19] Pearson, K.: On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**(11), 559–572 (1901). https://doi.org/10.1080/14786440109462720
[20] Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
[21] Schubert, E.: Stop using the elbow criterion for k-means and how to choose the number of clusters instead. ACM SIGKDD Explorations Newsletter **25**(1), 36–42 (2023)
[22] Trier, Ø.D., Jain, A.K., Taxt, T.: Feature extraction methods for character recognition – a survey. Pattern Recognition **29**(4), 641–662 (1996). https://doi.org/10.1016/0031-3203(95)00118-2
[23] Wahlberg, F., Wilkinson, T., Brun, A.: Historical manuscript production date estimation using deep convolutional neural networks. In: ICFHR. pp. 205–210 (2016). https://doi.org/10.1109/ICFHR.2016.0048
[24] West, G., Swindall, M.I., Brusuelas, J.H., Maltomini, F., Gerhardt, M., D'Angelo, M., Wallin, J.F.: A deep learning pipeline for the palaeographical dating of ancient Greek papyrus fragments. In: Pavlopoulos, John, e.a. (ed.) ML4AL ACL Workshop. pp. 177–185. ACL, Hybrid in Bangkok, Thailand and online (Aug 2024). https://doi.org/10.18653/v1/2024.ml4al-1.18, https://aclanthology.org/2024.ml4al-1.18/
[25] Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of CVPR. pp. 16133–16142 (2023)
[26] Zheng, M., Wang, F., You, S., Qian, C., Zhang, C., Wang, X., Xu, C.: Weakly supervised contrastive learning. In: ICCV. pp. 4009–4015 (2021), arXiv:2110.04770
[27] Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI. vol. 34, pp. 13001–13008 (2020)