Metric Learning for Ancient Coin Identification

Nathan R. Sprague*, Jason B. Forsyth†‡, Trevor Schonbrun†, Dhanshree D. Atre*, and Quinnie Lu†
*Department of Computer Science, †Department of Engineering, ‡Madison Art Collection
James Madison University
Email: spragunr@jmu.edu, forsy2jb@jmu.edu, schonbtk@dukes.jmu.edu, atre.dhanshree@gmail.com, luqx@dukes.jmu.edu

***Abstract*—Ancient coin identification is an instance-level recognition problem with a difficult data regime: large public collections provide many coin identities, but usually only one image per physical coin, while confirmed same-coin pairs from independent sources are scarce. Because such paired datasets are costly and likely to remain small, we ask how they can best be combined with abundant single-image public data. We test whether scarce paired data can be used effectively as an adaptation signal for tuning augmentations, which then provide synthetic repeated views of the public single-image identities used to train an embedding model. Our ArcFace-based pipeline tunes augmentation hyperparameters on repeated-image validation data and trains on public coin identities. For controlled evaluation, we introduce a curated dataset of 3,708 images covering 309 coin sides under varied lighting and reproduction conditions, including print–scan catalog-style artifacts. In this controlled setting, direct paired training achieves 98.01% Recall@1, while augmentation-guided public-data training achieves 94.57% without using paired identities as supervised training classes. Ablations identify augmentation tuning and preservation of intermediate spatial features as the principal contributors to public-data performance.**

***Index Terms*—metric learning, ArcFace, data augmentation, hyperparameter search, fine-grained recognition, ancient coin identification**

I. Introduction

When collecting and studying ancient artifacts, it is often important to establish an object's provenance, that is, its history of prior ownership. Provenance can affect an object's financial value, historical significance, authenticity, and even the legality of its ownership or sale [1]. In practice, however, many objects have incomplete or missing provenance records due to poor documentation, limited knowledge by collectors, or the relatively low value of an object compared with the effort required to research it [2]. Recovering this information often requires painstaking manual comparison against old invoices, auction listings, and catalog scans by trained experts. Because this process is labor intensive, provenance searches are often conducted only for especially valuable or unusual pieces. While commercial services such as L5 [3], ACSearch [4], and CoinArchives [5] offer automated provenance search, their catalog coverage and algorithms are proprietary, and none have disclosed accuracy on controlled test sets, making independent evaluation impossible.

This work is motivated by the goal of accelerating provenance recovery for ancient coins. In particular, we are interested in automatically comparing modern coin images against historical auction-catalog imagery in order to identify prior appearances of the same physical coin. Our longer-term application is the Madison Art Collection [6], which includes hundreds of ancient Greek and Roman coins together with a donated collection of auction catalogs that may contain useful provenance evidence. From a computer-vision perspective, this is an instance-level recognition problem: the task is not merely to recognize a coin type, but to determine whether two images depict the same individual specimen.

Image retrieval is a natural formulation for this task: a query coin image is compared against a gallery, and the system returns likely images of the same physical specimen for expert review. The central training challenge is the scarcity of paired supervision. Large public coin collections typically contain only a single image of each physical coin. The missing ingredient is verified repeated imagery of the same physical specimen from independent sources. Figure 1 illustrates the visual variation present in such pairs: a stater of Velia photographed for a 1972 print catalog and again in 2023 shows dramatic differences in reproduction quality and tonal rendering over five decades, while a Brutus denarius photographed by two different auction houses four years apart exhibits marked differences in lighting and color balance. Constructing even a few hundred such same-coin pairs would require substantial expert provenance work.

The setting has similarities to face recognition. Modern face-recognition systems are trained to produce embeddings that place images of the same identity near one another, and after training they can generalize to identities not seen during optimization. Ancient coin identification can be approached in a similar way, using metric learning to learn an embedding space rather than a closed set of coin labels. The analogy breaks down, however, in the availability of repeated identity data: successful face-recognition models are trained on large labeled datasets with many images per person, while any verified same-coin dataset from independent provenance sources is likely to be tiny. This makes the use of scarce paired data a central design question. In this work, we examine two boundary cases: using paired examples only as an adaptation signal to select augmentations for a model trained on much larger unpaired public collections, and using paired examples directly as supervised metric-learning data.

One motivation for the augmentation-guided strategy is to impose structure on how a tiny paired dataset can influence the learned embedding. With extremely limited paired supervision, overfitting to idiosyncrasies of the available examples is an inherent risk. A search over physically plausible image transformations acts as a prior over the kinds of variation expected in provenance imagery, such as changes in crop, lighting, tonal

rendering, and reproduction artifacts. This prior may reduce the risk of overfitting relative to fitting the embedding model directly on a tiny set of labeled repeated instances.

In this paper, we study how scarce paired data and abundant single-image public data can be combined in an ArcFace-based retrieval pipeline. Our main contributions are:

- **Controlled study of paired supervision.** We compare two strategies: training the embedding directly on repeated images, and using those images only to select an augmentation policy for an embedding trained on the larger public single-image dataset.
- **Multi-condition benchmark.** We introduce the JMU Sawhill multi-condition dataset, a dataset for evaluating instance-level retrieval under controlled lighting and physical print–scan variation.
- **Design findings.** Our ablations show that validation-guided augmentation tuning and preservation of intermediate spatial features account for most of the gains in public-data training, bringing it close to direct paired training in this controlled setting.

Source code for the full pipeline is available on GitHub.[1]

II. Related Work

A. Metric Learning and Face Recognition

Metric learning is a natural formulation when the goal is to recognize identities that may not have been seen during training. Rather than learning only a closed-set classifier, the model learns an embedding in which examples of the same identity are close and examples of different identities are separated. Early deep-learning face-recognition systems such as FaceNet helped establish this embedding-based formulation for verification, identification, and retrieval over identities that may be absent from the training set [7].

ArcFace is a widely used representative of this family of methods. It trains a normalized embedding with an additive angular margin, encouraging compact within-class features and greater angular separation between classes [8]. Although more recent face-recognition losses introduce adaptive or quality-aware margins [9], [10], we use ArcFace as a simple, well-established angular-margin baseline and focus on the effects of augmentation, spatial representation, and evaluation under the single-image-per-identity coin regime.

B. Ancient Coin Recognition

Computer vision analysis of ancient coins spans three problems of increasing difficulty: type identification (classifying a coin to a known numismatic type), die analysis (grouping coins by the die pair used to mint them), and instance recognition (determining whether two images show the same physical coin). Our work focuses on the instance recognition setting, where public datasets and reproducible evaluation protocols remain scarce.

[1]https://github.com/COIN-Research-Group/coin-embedding

Type identification is the best-studied problem, addressed with classical local feature matching [11], [12], learned keypoint configurations [13], and deep classifiers for coin motifs and portraiture [14]–[16], with recent work pushing toward open-set type recognition via Siamese Vision Transformers [17].

Die analysis sits at an intermediate granularity: because each ancient coin is struck from a unique die pair, grouping coins by die identity helps estimate issue sizes and study ancient economies [18]. Automated approaches include the Computer-Aided Die Study (CADS), a semi-automated tool that matches coins using ORB keypoints and proposes die groups via hierarchical clustering for a human reviewer to confirm [19]. Cornet et al. extend this pipeline by replacing the previously hand-engineered ORB features with XFeat, a learned deep keypoint detector and descriptor, and by automating the grouping step with graph-based label propagation instead of human review [20]. Fully unsupervised alternatives use Gaussian Process keypoints with Bayesian microclustering [21] and CNN features with k-means grouping [22]. These methods are effective within tightly controlled, single-type datasets. Each study examines coins sharing the same iconographic type, minting period, and general form factor, which substantially reduces the visual variation the system must handle.

Instance recognition is the focus of our paper: we ask whether two images depict the same physical coin. This problem is more challenging than identification or die studies as it requires distinguishing individual specimens across diverse types, cultures, and centuries, and under the uncontrolled photographic variation of real provenance imagery. The same coin can look strikingly different across the decades separating a historical catalog from a modern collection photograph (Figures 1a and 1b), and even across several years when photographed for a 2020 auction and a 2024 sale (Figures 1c and 1d). Commercial services such as L5, ACSearch, and CoinArchives offer automated provenance search, but their catalogs and algorithms are proprietary, making independent evaluation impossible [3]–[5]. To our knowledge, our dataset and pipeline provide the first public reproducible benchmark focused specifically on instance-level ancient coin retrieval.

C. Data Augmentation and Augmentation Search

Automatic augmentation methods search for transformation policies that improve validation performance, rather than relying on a single hand-selected preprocessing pipeline. AutoAugment introduced this train-and-select pattern with a learned controller over discrete image transformations [23]; later methods reduce the search cost or simplify the parameterization through density matching, differentiable relaxation, global magnitude controls, or randomized single-operation policies [24]–[27]. Our use of augmentation search differs in both objective and data regime: rather than tuning for closed-set classification accuracy, we tune domain-specific augmentation parameters for retrieval accuracy on repeated images of the same physical coins. Because each public training identity has only one observed image, the augmentation distribution

(a) Stater of Velia from 1972 Sale of President John Q. Adams and Descendants

(b) Same coin as (a) but photographed in 2023. White labeling has been added to coin edge for identification in a collection.

(c) Denarius of M. Junius Brutus, Harlan J. Berk BBS Sale 227, Lot 443, March 2024.

(d) Same coin as (c) but photographed by Heritage Auctions in November 2020.

Fig. 1: Comparison of coins under different photography and cataloging conditions

supplies the within-identity variation used to train the metric model. Thus our procedure keeps AutoAugment's broad train-and-select structure, but replaces the learned controller with black-box optimization over domain-specific parameters and replaces classification validation accuracy with a retrieval-based validation score.

D. Spatial Invariance in Convolutional Networks

Standard classification backbones become increasingly invariant to position, scale, and local deformation as depth increases. This invariance is useful for category recognition, but it can discard spatial evidence that matters for instance-level retrieval. Prior work on hypercolumns, hierarchical metric learning, and local/global retrieval models similarly shows that intermediate or localized features can improve matching when fine spatial detail remains discriminative [28]–[31]. We therefore evaluate a lightweight alternative to specialized retrieval architectures: truncated backbones that remove the final pooling stages and train an embedding head on the retained spatial feature map. This tests whether preserving a higher-resolution spatial representation helps coin retrieval, where discriminative evidence may lie in relief, wear, inscriptions, and flan shape.

III. Dataset

Our experimental setup uses two distinct data sources with complementary roles. Large public coin datasets supply the unpaired training identities needed to learn a general embedding space. The JMU Sawhill multi-condition dataset supplies repeated images of the same physical coins under varied acquisition conditions. We use this paired dataset in two ways: as a small validation and test resource for augmentation-guided public-data training, and as direct supervised training data for a simpler paired-training baseline.

A. Public Training Data

We train primarily on two public ancient-coin datasets: RRC-60 [32] and RRCD [15]. RRC-60 contains 12,000 images of Roman Republican coins from 60 coin types, with 100 obverse images and 100 reverse images for each type. RRCD contains more than 18,000 images spanning 228 reverse motifs. These datasets provide substantial visual diversity for training, but they are organized for type or motif classification rather than verified same-coin retrieval; they do not provide repeated independent images of the same physical coin.

We combine the RRC-60 and RRCD images into a single public-data pool and split it into 20,671 training images, 2,953 validation images, and 5,907 held-out test images, corresponding to approximately 70%, 10%, and 20% of the public data, respectively. The public validation split supports ordinary hyperparameter and checkpoint selection. The held-out public test split is not used to train or tune the embedding model; instead, it is used as a pool of public-collection distractors in the paired retrieval evaluations.

B. The JMU Sawhill multi-condition dataset

Beyond the publicly available datasets, we developed a novel dataset using a variety of lighting conditions to simulate photographic variation that might be encountered when locating a coin in historical references. To generate this set, we selected 155 coins from the Madison Art Collection. The collection contains over 400 ancient Greek and Roman coins assembled by Dr. John Sawhill from the 1930s until the 1970s [6].

Our dataset spans a much wider historical and geographic range than RRCD and RRC-60. Whereas those datasets contain only Roman Republican coinage from approximately 137 BCE to 49 BCE, our dataset spans ca. 540 BCE to 120 CE, including 46 Greek city-states and kingdoms, 24 Roman rulers and officials, and 25 distinct denominations struck in silver, bronze, and gold. This wide geographic and temporal span produces significant variation in coin design, size, and material.

From this collection, 85 Roman and 69 Greek coins were selected for two-sided imaging. Each coin was placed on an automated turntable inside a light box with a strong directional light; video was recorded from an overhead camera as the coin rotated through 90 degrees in each direction, for both the obverse and reverse sides. One additional coin was imaged only on the reverse. Twelve frames were extracted from each side's video, yielding 309 imaged coin sides and 3,708 images in total.

Paired data are split by physical coin, with both sides kept together. The 109-side development partition (1,308 images) supports augmentation selection and direct training. Another 200 sides (2,400 images) are held out for testing.

C. Reproduction-Variation Pipeline

After the video-based acquisition described above, we converted the extracted stills into the final evaluation images using a post-collection reproduction-variation pipeline. The purpose of this stage was to simulate the appearance changes that occur when a coin image is copied, printed, scanned, and recropped. The pipeline began with the twelve extracted stills for each coin side and replaced them with a fixed set of digital and print–scan variants while preserving the identity of the underlying physical coin.

Each coin side yielded twelve reproduction variants: four fully digital variants, consisting of two color and two grayscale images, and eight print–scan variants. The twelve source stills were randomly assigned one-to-one to the twelve variants, after which each received a mild random in-plane rotation and scale change. For each print–scan variant, the transformed image was then placed on a printed sheet.

The eight print–scan variants crossed print–scan station, color mode, and printed size. We used two print–scan stations, each with one scanner and two physical printers: one color printer and one grayscale printer. At each station, coins were printed in color and grayscale at two sizes, approximately 1.2 inches and 1.8 inches across.

After the printed sheets were scanned, each coin was cropped, centered on a square canvas, and resized to 512×512 pixels. The final result is that each coin side has twelve output images: four direct digital images and eight print–scan images spanning two print–scan stations, four printers, two color treatments, and two printed scales. Figure 2 shows the complete set of variants for one coin side.

The entire dataset of original images and the reproduction variations is available online[2].

[2]https://huggingface.co/datasets/COIN-Research-Group/sawhill-dataset

IV. Method

A. ArcFace-Based Coin Identification

We formulate coin identification as metric learning over physical coin identities. Let $f_\theta(x) \in \mathbb{R}^d$ denote the embedding produced by a convolutional backbone and embedding head for an input coin image x. In the augmentation-guided public-data setting, each public-dataset coin image is treated as a separate identity class. The training set therefore supplies many distinct coin identities, but only one observed view of each identity; additional training views are generated through augmentation rather than through repeated real photographs of the same physical coin.

For each training image, the dataset loader applies the selected augmentation pipeline on the fly and assigns the augmented sample the label of its source image. The model is trained with ArcFace, an additive angular-margin softmax loss over normalized embeddings and class weights [8]. For an embedding $z_i = f_\theta(x_i)$ with class label y_i, ArcFace increases the angular separation required for the correct class while keeping the optimization in a standard multiclass-classification form:

$$\mathcal{L}_{\text{ArcFace}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{j\neq y_i} e^{s\cos\theta_j}}, \quad (1)$$

where θ_j is the angle between the normalized embedding and class weight for class j, m is the angular margin, and s is a scale factor.

At inference time, the ArcFace classification layer is discarded and only f_θ is used. Query and gallery images are embedded independently, their embeddings are ℓ_2-normalized, and matches are ranked by cosine similarity.

All trained ArcFace models use 224×224 inputs and 512-dimensional embeddings and are optimized with Adam. The ArcFace loss uses the default margin $m = 28.6^\circ$ and scale $s = 64$. The headline direct and public-data models both use truncated ConvNeXt-Tiny, a batch size of 384, and a learning rate of 10^{-4}. Complete training configurations and augmentation-search bounds are provided in the released code.

B. Augmentation Policy

Training uses an on-the-fly stochastic augmentation policy to turn each single public coin image into a stream of perturbed views. Each policy is a composition of geometric, photometric, and reproduction-inspired transforms: random resized crop, random affine rotation and translation, random perspective distortion, color jitter, Gaussian blur, random grayscale conversion, and, in lighting-enabled variants, directional relighting.

We compare two ways of setting the parameters of this policy. The first uses a hand-selected default policy, chosen from visual inspection and preliminary training runs. The second uses a tuned policy whose parameters are selected by the validation-guided search described below. In both cases, once a policy has been selected, training samples random transformations from that fixed policy distribution.

Fig. 2: Example outputs from the reproduction-variation pipeline for a single coin side. Note the significant lighting variations and reproduction artifacts.

As a coin-specific augmentation operator, we introduce directional relighting to address the strong dependence of the coin's surface detail appearance on illumination direction. The operator uses the image itself to approximate a coin surface elevation field from intensity gradients at multiple spatial scales, treating brighter local structure as higher relief. These gradients define an approximate normal field, which is shaded with a virtual directional light using a Lambertian model.

For a fixed augmentation policy ϕ, we train the embedding model with sampled transformations from that distribution:

$$\theta^*(\phi) = \arg\min_{\theta} \ \mathcal{L}_{\text{train}}(\theta, \phi), \tag{2}$$

where $\mathcal{L}_{\text{train}}$ is the ArcFace training loss on the public coin datasets.

C. Augmentation Hyperparameter Search

Because augmentation is the only mechanism by which the model observes within-identity variation during training, we select augmentation settings with black-box augmentation-policy search. We separate this search from ordinary training-hyperparameter selection: learning rate and batch size are chosen with a coarse grid search, then held fixed while the augmentation policy is tuned. To support the lighting ablation, we run this augmentation search twice: once over the base policy without directional relighting and once over the lighting-enabled policy.

The search variables include the magnitudes of the operators in the policy and, where applicable, probabilities for stochastic choices such as grayscale conversion and directional relighting. In the lighting-enabled search, the space also includes directional-lighting parameters such as light direction, strength, and ambient floor. The search domains are bounded to produce plausible catalog-style variation, with exact bounds provided in the released implementation.

Each augmentation-search trial trains a model with candidate augmentation parameters ϕ on the public single-image training set and scores the resulting embeddings on the Sawhill repeated-image validation split:

$$\phi^* = \arg\max_{\phi} \ \text{score}_{\text{val}}(\theta_\phi), \tag{3}$$

where θ_ϕ denotes the public-data model trained with augmentation settings ϕ. The validation score measures retrieval of the same coin sides across different capture and reproduction conditions, rather than closed-set classification accuracy. Thus the search favors augmentation settings that improve invariance to acquisition changes represented in the paired validation set, while leaving those paired identities out of the public-data training objective.

We use Optuna's Tree-structured Parzen Estimator (TPE) as the practical optimizer for this augmentation-policy search [33], [34]. Optuna proposes candidate augmentation settings; each candidate is evaluated by training a new embedding model with the fixed training hyperparameters, and the best setting is fixed before final evaluation on the held-out test split. Each TPE search used 250 trials, with each trial model trained for 15 epochs. This budget allows enough training for a meaningful retrieval signal while keeping the search computationally feasible.

D. Direct Paired-Training Baseline

The augmentation-guided approach uses the paired data to select an augmentation policy for public-data training. For comparison, we also evaluate the alternative allocation of paired supervision: training directly on paired reproduction variants from Sawhill. In this baseline, each coin side in Sawhill is treated as an identity class, and its repeated variants provide supervised within-identity examples. The model architecture, ArcFace objective, and retrieval procedure are otherwise unchanged.

E. Truncated Backbone Models

In addition to standard pooled CNN backbones, we evaluate truncated variants that preserve an intermediate spatial feature

map. The motivation is that coin images are typically presented in a roughly canonical pose, and fine local correspondences in the relief, inscriptions, flan shape, and wear patterns may be useful for instance-level matching. Standard image-classification backbones deliberately remove much of this spatial information through the final stages of the network and global average pooling. Our truncated models therefore stop earlier and use the retained spatial tensor as the input to a learned embedding head.

We implemented this truncation design for both ResNet-50 [35] and ConvNeXt-Tiny [36] during development; the reported learned models use ConvNeXt-Tiny. For ConvNeXt-Tiny, the truncated model keeps the feature extractor through the stage that outputs a 14×14 feature map with 384 channels, discarding the later stages, global pooling, layer-normalization classifier block, and ImageNet classifier. A 1×1 convolution reduces the feature map to 32 channels, after which the flattened spatial tensor is projected by the embedding head before being passed to ArcFace.

The ArcFace objective is otherwise unchanged: each architecture produces an embedding vector, and the ArcFace classification layer is parameterized with the corresponding embedding dimensionality. Thus the truncation changes only the backbone-to-embedding mapping, not the metric-learning loss or retrieval procedure.

V. Experimental Design

A. Evaluation Protocol and Metrics

We evaluate on held-out coin identities from Sawhill, with held-out public-data images included as additional distractors. The primary evaluation is a single-match retrieval task. For each query image from the held-out test set, the gallery contains exactly one known matching image of the same coin side and a set of distractors. The full search database is fixed across every reported result: it comprises all 2,400 held-out reproduction images from Sawhill (200 coin sides $\times$ 12 variants) together with 5,907 held-out public-data images, for a total of 8,307 images. Distractors therefore include both within-collection confusers and broader public-collection distractors. This setting models a provenance-search workflow in which a system returns a short ranked list for expert review. We report Recall@1, Recall@5, and Recall@10; Recall@1 measures whether the correct match is the top-ranked result, while Recall@5 and Recall@10 reflect the usefulness of a short candidate list. For each coin side, we evaluate all 12×11 query–target pairs.

As a secondary evaluation, we use a multi-match retrieval task. For each query image from Sawhill, the gallery contains the 11 known matching images for that coin, excluding the query image itself if it appears in the gallery, together with the same distractor pool. This setting measures whether the embedding space ranks multiple views of the same coin side ahead of non-matching coins, rather than testing only whether a single target image is recovered. For this task, we report mean average precision (mAP), which accounts for the ranks of all relevant images for each query.

B. Comparison Strategy

Our comparisons are organized around how much task-specific training is used and how scarce paired data enters the learning process. Frozen ConvNeXt and DINOv2 [37] features provide reference points for retrieval without coin-specific metric learning. Public-data ArcFace models train the embedding on large single-image coin datasets, with paired data from Sawhill used only for augmentation-policy selection. Direct paired training represents the alternative allocation of scarce paired data, using reproduction variants from Sawhill directly as supervised identity classes.

Within the public-data ArcFace setting, we further ablate the augmentation policy, lighting augmentation, and backbone truncation to identify which design choices account for the best public-data result.

VI. Results

We report results in two stages. Table I compares the main training and representation regimes. Table II then ablates the public-data ArcFace model to separate the effects of augmentation tuning, lighting augmentation, and backbone truncation.

The main comparison in Table I shows that generic frozen features are not enough for this instance-level task, even when the underlying representation is strong. With both regimes evaluated on the same gallery and the same truncated backbone, direct paired training gives the strongest result, at 98.01% Recall@1. Public-data training with tuned augmentation comes within roughly three points (94.57%) despite never using paired coin identities as supervised classes.

Table II identifies augmentation tuning and backbone truncation as complementary sources of the performance gain. Tuning raises Recall@1 from 49.92% to 76.97%, while truncation with manual augmentation raises it to 71.01%. Combining them reaches 94.17%, 17.20 points above tuning alone and 23.16 points above truncation alone. Lighting augmentation is inconsistent: it helps the truncated model with manual augmentation (+4.96 points) but is negligible or slightly harmful elsewhere, adding just 0.40 points to the best model and reducing the tuned standard model. Once tuning and truncation are in place, it offers no reliable benefit.

VII. Limitations

The main limitation of this study is the absence of a fully natural repeated-image dataset for ancient coin instance identification. The ideal evaluation set would contain multiple independently acquired images of the same physical coins from different auction houses, catalogs, photographers, scanners, and time periods. Such data are difficult to obtain at scale because the identity links between historical images and modern coin photographs are often precisely the unknown provenance evidence being sought. The JMU Sawhill multi-condition dataset provides controlled repeated images and print–scan reproductions, but it cannot fully capture the diversity and messiness of real provenance-search imagery.

This controlled setting affects the interpretation of both direct paired training and augmentation-guided public-data

TABLE I: Retrieval results comparing the main training and representation regimes.

Method	Single-match retrieval			Multi-match retrieval
	Recall@1 ↑	Recall@5 ↑	Recall@10 ↑	mAP ↑
Frozen ConvNeXt features	15.63%	24.75%	30.18%	0.2824
DINOv2 (frozen features)	37.30%	52.42%	60.04%	0.5723
Direct paired ArcFace training	98.01%	99.30%	99.62%	0.9942
Public-data ArcFace, tuned aug. + trunc. + lighting	94.57%	97.66%	98.34%	0.9800

TABLE II: Ablation study for public-data ArcFace training.

Configuration	TPE Aug.	Lighting Aug.	Trunc.	Single-match retrieval			Multi-match retrieval
				R@1 ↑	R@5 ↑	R@10 ↑	mAP ↑
Standard ConvNeXt	–	–	–	49.92%	63.83%	69.81%	0.6767
Standard ConvNeXt + TPE-tuned aug.	✓	–	–	76.97%	85.83%	88.84%	0.8774
Standard ConvNeXt + lighting aug.	–	✓	–	50.00%	64.08%	69.84%	0.6762
Truncated ConvNeXt + manual aug.	–	–	✓	71.01%	82.24%	86.22%	0.8482
Standard ConvNeXt + TPE-tuned + lighting aug.	✓	✓	–	72.11%	81.56%	85.28%	0.8397
Truncated ConvNeXt + TPE-tuned aug.	✓	–	✓	94.17%	97.36%	98.29%	0.9782
Truncated ConvNeXt + lighting aug.	–	✓	✓	75.97%	86.53%	90.14%	0.8865
Truncated ConvNeXt + TPE-tuned + lighting aug.	✓	✓	✓	94.57%	97.66%	98.34%	0.9800

training. Direct paired training may benefit from repeated examples drawn from the same acquisition and reproduction process used at evaluation time, while the tuned augmentation policies are also selected against this controlled proxy distribution. The comparison is therefore best understood as a controlled study of how scarce paired supervision can be allocated, rather than as a final estimate of performance under fully independent catalog-to-catalog variation. Because we report one run per experimental configuration, we do not claim statistical significance for small differences between methods.

A second limitation is the cost of augmentation-policy search. Each TPE trial requires training an embedding model and evaluating retrieval performance on repeated-image validation data. This makes the search computationally expensive and limits how easily the full procedure can be repeated for a new institution, collection, scanner, catalog source, or target image distribution.

Finally, the current system treats each coin side as a separate identity. This simplifies training and evaluation, and it matches common image-level retrieval workflows, but it ignores the fact that real coins are inherently two-sided objects. Combining evidence from obverse and reverse images, when both are available, could improve identification performance and reduce false matches. Multi-side retrieval and score fusion are therefore natural extensions of the present work.

VIII. Conclusion

We introduced the JMU Sawhill multi-condition dataset, a controlled multi-condition benchmark for instance-level ancient coin retrieval, and compared two uses of scarce paired supervision: direct training on repeated-image identities and augmentation selection for a model trained on public single-image identities. The benchmark includes controlled lighting changes and print–scan reproduction artifacts. On this controlled proxy, direct paired training produced the strongest retrieval results, while augmentation-guided public-data training remained competitive without using paired identities as supervised training classes.

A natural next step is to move beyond these two boundary cases by combining real paired examples with augmentation-generated positives during training. Rather than using paired data only to select augmentations, or using only real paired variants as the source of within-identity variation, a joint training regime could sample both kinds of positive evidence in the metric-learning objective. Different weights on real repeated images and synthetic augmented views would then control how strongly the model trusts scarce verified pairs relative to the broader augmentation distribution.

More broadly, this work frames ancient coin provenance search as a problem not only of choosing a metric-learning architecture, but also of determining how scarce same-instance supervision should be allocated. The JMU Sawhill multi-condition dataset provides a reproducible testbed for studying that question under controlled acquisition and reproduction variation.

Acknowledgment

Generative AI tools were used in the development of source code for experiments and in the preparation of portions of the manuscript text. The authors take full responsibility for all content in this paper, including the accuracy, originality, and integrity of any AI-assisted material.

All photos are used with permission or under fair use.

References

[1] UNESCO, "Convention on the means of prohibiting and preventing the illicit import, export and transfer of ownership of cultural property," 1970, adopted 14 November 1970.

[2] N. Chipangura, Y. Cai, and S. Delepierre, "Provenance research in museums: Principles, practices and possibilities," *Museum International*, vol. 77, no. 3–4, 2025.

[3] L5, "L5 Provenance Search," 2026, accessed: 2026-05-20. [Online]. Available: https://l5.com/Provenance/Index.aspx

[4] ACSearch, "acsearch.info," 2026, accessed: 2026-05-20. [Online]. Available: https://www.acsearch.info/

[5] CoinArchives, "Coinarchives," 2026, accessed: 2026-05-20. [Online]. Available: https://www.coinarchives.com/

[6] James Madison University, Madison Art Collection, "The Sawhill collection," 2026, accessed: 2026-05-20. [Online]. Available: https://www.jmu.edu/madisonart/_sawhill_collection.shtml

[7] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[9] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 5900–5909.

[10] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality Adaptive Margin for Face Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 18 750–18 759.

[11] M. Kampel and M. Zaharieva, "Recognizing ancient coins based on local features," in *Advances in Visual Computing (ISVC 2008)*, ser. Lecture Notes in Computer Science, 2008, pp. 11–22.

[12] O. Arandjelović, "Reading ancient coins: Automatically identifying Denarii using obverse legend seeded retrieval," in *European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, vol. 7575. Springer, 2012, pp. 317–330.

[13] O. Arandjelović, "Automatic attribution of ancient Roman imperial coins," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2010, pp. 1728–1734.

[14] I. Schlag and O. Arandjelović, "Ancient Roman coin recognition in the wild using deep learning based recognition of artistically depicted face profiles," in *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 2898–2906.

[15] H. Anwar, S. Anwar, S. Zambanini, and F. Porikli, "Deep ancient Roman Republican coin classification via feature fusion and attention," *Pattern Recognition*, vol. 114, p. 107871, Jun. 2021.

[16] C. Kiourt and V. Evangelidis, "AnCoins: Image-Based Automated Identification of Ancient Coins Through Transfer Learning Approaches," in *Pattern Recognition. ICPR International Workshops and Challenges*, Cham, 2021, pp. 54–67.

[17] Z. Guo, O. Arandjelović, D. Reid, Y. Lei, and J. Büttner, "A Siamese Transformer network for zero-shot ancient coin classification," *Journal of Imaging*, vol. 9, no. 6, p. 107, 2023.

[18] C. M. Kraay, *Archaic and Classical Greek Coins*. London: Methuen, 1976.

[19] Z. M. Taylor, "The computer-aided die study (CADS): A tool for conducting numismatic die studies with computer vision and hierarchical clustering," Honors Thesis, Trinity University, 2020, Computer Science Honors Theses, no. 54.

[20] C. Cornet, H. Aumaitre, R. Besançon, J. Olivier, T. Faucher, and H. Le Borgne, "Automatic die studies for ancient numismatics," in *Computer Vision – ECCV 2024 Workshops*, ser. Lecture Notes in Computer Science, 2025, pp. 246–262.

[21] A. Heinecke, E. Mayer, A. Natarajan, and Y. Jung, "Unsupervised statistical learning for die analysis in ancient numismatics," *arXiv preprint arXiv:2112.00290*, 2021.

[22] C. Deligio, C. von Nicolai, M. Möller, K. Rösler, J. Tietz, R. Krause, K. P. Hofmann, K. Tolle, and D. Wigg-Wolf, "IT- and machine learning-based methods of classification: The cooperative project ClaReNet – Classification and Representation for Networks," *Zenodo*, 2022.

[23] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning Augmentation Strategies From Data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 113–123.

[24] S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim, "Fast AutoAugment," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[25] Y. Li, G. Hu, Y. Wang, T. M. Hospedales, N. M. Robertson, and Y. Yang, "DADA: differentiable automatic data augmentation," in *The European Conference on Computer Vision (ECCV)*, 2020.

[26] E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le, "RandAugment: Practical Automated Data Augmentation with a Reduced Search Space," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 18 613–18 624.

[27] S. G. Müller and F. Hutter, "TrivialAugment: Tuning-Free Yet State-of-the-Art Data Augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[28] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, "Hypercolumns for Object Segmentation and Fine-Grained Localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 447–456.

[29] M. E. Fathy, Q.-H. Tran, M. Z. Zia, P. Vernaza, and M. Chandraker, "Hierarchical Metric Learning and Matching for 2D and 3D Geometric Correspondences," in *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, Berlin, Heidelberg, Sep. 2018, pp. 832–850.

[30] B. Cao, A. Araujo, and J. Sim, "Unifying Deep Local and Global Features for Image Search," in *Computer Vision – ECCV 2020*, Cham, 2020, vol. 12365, pp. 726–743.

[31] M. Yang, D. He, M. Fan, B. Shi, X. Xue, F. Li, E. Ding, and J. Huang, "DOLG: Single-Stage Image Retrieval With Deep Orthogonal Fusion of Local and Global Features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 772–11 781.

[32] S. Aslan, S. Vascon, and M. Pelillo, "Two sides of the same coin: Improved ancient coin classification using graph transduction games," *Pattern Recognition Letters*, vol. 131, pp. 158–165, 2020.

[33] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA, Jul. 2019, pp. 2623–2631.

[34] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for Hyper-Parameter Optimization," in *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[36] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 11 976–11 986.

[37] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024.