JOURNAL OF BIOMEDICAL RESEARCH & ENVIRONMENTAL SCIENCES
issn: 2766-2276
SUBJECT AREA(S): ARTIFICIAL INTELLIGENCE
DOI: 10.37871

SHORT COMMUNICATION

Style-Aware Hierarchical Recognition for Accessible Museum Guidance Systems

Youngbok Cho*

Department of Computer Education, Gyeongkuk National University, Andong, Republic of Korea

Abstract

Background: Museum artwork recognition presents distinct challenges: extreme data scarcity (typically one training image per title), museum-specific imaging conditions, and high intra-class visual variability that requires sensitivity to artistic style rather than semantic content alone.

Methods: We present a Style-Aware Hierarchical Recognition Framework that addresses these challenges through three coordinated components. First, a domain-specific augmentation pipeline generates ten synthetic variants per original image, simulating spot lighting, lens distortion, motion blur, and partial occlusion to enrich training diversity. Second, the Style-Aware Feature Fusion (SAFF) module is mounted atop an Inception ResNetV2 backbone; it extracts a 256-dimensional global style embedding via average pooling, generates channel-wise attention weights through a sigmoid-activated dense layer, and reweights the multi-scale feature map in a residual manner to emphasize stylistically discriminative cues such as brushstroke texture, color layering, and compositional rhythm. Third, a two-stage hierarchical classifier first assigns each artwork to one of four departments, then applies a department-specific title-level classifier, reducing the effective classification space from 261 to 42–86 classes per model.

Results: On the Metropolitan Museum of Art benchmark, the framework achieves 86% department-level accuracy (F1 = 0.86) and 84% weighted-average Top-1 accuracy across 261 title classes (F1 = 0.81). The end-to-end hierarchical pipeline yields 75% final Top-1 accuracy, substantially outperforming flat baselines including InceptionResNetV2 alone (56%), InceptionV3 (45%), and VGG16 (37%). Ablation analysis confirms that the style embedding is the dominant performance factor (−9% when removed), and SAFF outperforms established attention mechanisms (SE-Net, CBAM, and transformer self-attention) under identical training conditions.

Conclusion: Combining domain-aware augmentation, style-sensitive feature fusion, and hierarchical decomposition provides a robust and scalable foundation for museum artwork recognition under extreme data constraints, with direct applicability to accessible museum guidance systems.

***Corresponding author(s)**

Youngbok Cho, Department of Computer Education, Gyeongkuk National University, Andong, Republic of Korea
Email: ybcho@gknu.ac.kr
DOI: 10.37871/jbres2319

Submitted: 14 July 2026

Accepted: 28 July 2026

Published: 31 July 2026

Keywords

- Artwork classification
- Domain-specific augmentation
- Style-Aware Feature Fusion (SAFF)
- InceptionResNetV2
- Hierarchical classification
- Few-shot learning

How to cite this article: Cho Y. Style-Aware Hierarchical Recognition for Accessible Museum Guidance Systems. J Biomed Res Environ Sci. 2026 July 31; 7(7): 13. Doi: 10.37872/jbres2319

Abbreviations

SAFF: Style-Aware Feature Fusion; CNN: Convolutional Neural Network; OCR: Optical Character Recognition; t-SNE: t-Distributed Stochastic Neighbor Embedding; SE-Net: Squeeze-and-Excitation Network; CBAM: Convolutional Block Attention Module; FLOPs: Floating Point Operations; INT8: 8-bit Integer Quantization; GPU: Graphics Processing Unit; RGB: Red-Green-Blue (image channel format)

Introduction

Art museums serve as repositories of cultural heritage and spaces for emotional and intellectual engagement. Appreciating an artwork depends on access to contextual information, including the title, the artist, the historical period, and the interpretive meaning. Visitors with visual impairments cannot access this information visually; they rely on smartphone cameras to examine details, revealing the fundamental limitation of visual access without contextual identification. To bridge this gap, an assistive system must recognize each artwork before providing a relevant, meaningful description.

Existing solutions, such as audio guides or QR-code-based applications, require either manual input or preinstalled infrastructure, limiting their spontaneity and scalability. These systems also fail to support visitors who desire real-time, seamless interaction in front of any artwork, particularly in changing or unstructured exhibition layouts. Smartphone-based artwork recognition offers a compelling alternative: by capturing a photograph, users can instantly retrieve an artwork's title and contextual interpretation, enabling autonomous, personalized museum experiences.

While accessibility is the primary motivation, smartphone-based recognition also serves the general public. Casual visitors often lack the background knowledge to engage deeply with artworks and hesitate to use unfamiliar interfaces. A recognition-based system democratizes access to information and supports broader applications such as digital archiving, metadata tagging, and personalized museum guides.

Convolutional Neural Networks (CNNs) have revolutionized computer vision, achieving human-level performance on image classification [1,2]. However, off-the-shelf models struggle with the fine-grained, high intra-class variability of museum collections, where each artwork is often represented by a single high-resolution image captured under complex, museum-specific imaging conditions.

In a museum setting, image recognition faces a unique confluence of challenges. Illumination variance from spotlights or glass glare, off-axis viewpoints from handheld devices, and textural inconsistencies caused by aging surfaces or reflections can severely degrade recognition accuracy. Moreover, with only a single exemplar image per title, training data for each class is extremely limited, aggravating overfitting and making traditional flat classifiers impractical.

Artwork recognition is not merely an object recognition task — it demands perception of artistic style. Paintings with similar subjects differ entirely in brushstroke technique, surface texture, color layering, and compositional rhythm. Generic CNNs, trained on natural-scene images, overlook these stylistic nuances and focus on semantic content rather than material or technique. Style-sensitive feature extraction is therefore essential: the model must attend to subtle cues such as impasto depth, line quality, or varnish gloss that distinguish one artist's rendition from another.

Directly classifying among hundreds of unique titles poses two major bottlenecks. First, with only seven training images per title on average, the scenario is essentially a few-shot learning problem that substantially increases overfitting risk. Second, a softmax output layer

over hundreds of classes demands excessive memory and computation, hindering scalability. These issues motivate a hierarchical recognition strategy in which each artwork is first assigned to a coarse department before title-level classification, reducing the downstream classification space and mitigating the few-shot learning challenge.

The main novelty of the proposed framework is threefold and distinguishes it clearly from prior artwork recognition methods. Existing approaches either (i) apply generic deep learning pipelines that treat artwork recognition as standard object classification [3,4] — ignoring the stylistic rather than semantic nature of the discrimination task — or (ii) use attention mechanisms such as SE-Net or CBAM that were designed for natural-scene recognition and lack an explicit global style representation or metric-learning objective [5]. Unlike these methods, the proposed Style-Aware Feature Fusion (SAFF) module is the first attention mechanism designed specifically for fine-grained artwork discrimination: it extracts a 256-dimensional global style embedding *via* average pooling and trains it jointly with a triplet metric loss, forcing the embedding space to be stylistically coherent. Combined with a hierarchical two-stage classification strategy — which reduces the per-model classification space from 261 to 42–86 classes and directly addresses the extreme one-shot-per-title data constraint — and a museum-tailored augmentation pipeline, the framework is the first to integrate all three components in a unified architecture purpose-built for the few-shot museum artwork recognition setting.

Thus, style-aware modeling and hierarchical decomposition are not optional design choices but core necessities for reliable, efficient, and scalable artwork recognition in real-world museum environments. Our key contributions are:

(1) A domain-specific augmentation pipeline tailored to museum imagery, generating ten synthetic variants per original image to simulate spot lighting, lens distortion, motion blur, and partial occlusion.

(2) The Style-Aware Feature Fusion (SAFF) module, which integrates style embeddings with CNN feature maps in a lightweight residual fashion. SAFF extracts a global style vector *via* average pooling, generates channel-wise attention weights through a sigmoid-activated dense layer, and reweights the multi-scale feature map to emphasize both low-level (texture, color) and high-level (composition, layout) stylistic cues, encouraging the model to focus on artist-specific patterns critical for distinguishing visually similar artworks.

(3) A two-stage hierarchical training strategy that preserves global style context while reducing per-model class counts, yielding faster convergence and improved robustness on rare classes.

Related Work

Assistive vision systems

Smartphone-based assistive technologies have rapidly evolved to support users with visual impairments in daily activities. Early surveys highlight the potential of off-the-shelf mobile devices, documenting features such as OCR-based text reading and object detection to compensate for declining vision, especially in low- and middle-income settings [6].

A systematic review of over 50 smartphone applications designed to assist users with various daily tasks found that these applications range from screen reading to environmental sensing, leveraging camera input, artificial intelligence, and haptic feedback for navigation support, object identification, and reading assistance [7]. Volunteer-driven platforms such as "Be My Eyes" connect blind users *via* live video to sighted helpers for ad hoc assistance, demonstrating high user satisfaction, although such systems

rely on real-time human involvement, which causes delays and limits scalability [8].

Domain-specific museum solutions are also emerging. The Houston Museum of Natural Science's "ReBokeh" application applies real-time contrast and zoom filters to enhance low-vision viewing of exhibits [9]. AI-powered wearable systems such as "Ask Envision" integrate large-scale vision models to offer contextual descriptions through smart glasses, achieving more seamless interaction while raising concerns over misidentification in safety-critical scenarios [10].

Artwork classification techniques

The computer vision community has explored fine-grained artwork recognition extensively. Early CNN efforts focused on artist attribution: Balakrishnan, et al. [3] trained a standard CNN on Rijksmuseum images to classify painters, illustrating the feasibility of deep features for capturing style cues. Subsequent work employed custom CNN architectures for painter identification, reporting classification rates above 80% on curated painting subsets [4]. Researchers have also tackled genre and medium recognition; Johnson and Lee [11] applied a VGG-style network to distinguish Impressionist from Baroque works, achieving over 75% genre accuracy.

Recent advances combine CNNs with attention mechanisms and transformer modules. Wang and Song [5] proposed a fusion model integrating locally extracted CNN features with global context *via* a transformer backbone, yielding a 9–10 percentage point improvement in classification accuracy on Chinese and oil-painting datasets. Despite these successes, most prior work assumes abundant samples per class, whereas museum collections offer only a single exemplar per title — a constraint that motivates the domain-specific augmentation and hierarchical strategy proposed in this work.

More recent work has explored few-shot classification methods for artwork recognition. Elkholy, et al. [12] applied Prototypical Networks to the WikiArt dataset, reporting Top-1 accuracies between 55% and 63% under 5-shot conditions — comparable to our metric-learning baselines but below our hierarchical framework's 75%. Condoleo, et al. [13] fine-tuned a Vision Transformer (ViT-B/16) on the SemArt dataset for artwork classification, achieving 71% accuracy when trained with full supervision; however, their setting assumed far more training images per class (hundreds rather than one). In the few-shot domain, Yanagi, et al. [14] demonstrated that style-specific prototype augmentation improves Prototypical Network performance on painting classification by up to 8 percentage points, supporting the importance of style-aware feature design — consistent with our SAFF ablation findings. Compared with all these recent approaches, our framework uniquely combines a purpose-built style embedding, a domain-specific museum augmentation pipeline, and hierarchical classification in a single architecture that operates from a single training image per title.

Methods

Dataset and Augmentation pipeline

We employed an end-to-end augmentation pipeline that generates ten synthetic variants for each class from a single original image, followed by a 60:20:20 train/validation/test split. All operations are orchestrated in Python using Albumentations [15] for transformation composition and Python's glob module for flexible file discovery based on filename patterns.

Artwork Data Acquisition: We queried the Metropolitan Museum of Art's (Met) public BigQuery table to retrieve artworks marked as "highlight." Among these, we selected works classified as Prints, Drawings, or Paintings and

extracted key metadata fields including title, department, and artist name. For each artwork, a headless browser visited its detail page to extract the image URL and downloaded the image into a folder organized by culture. Metadata, including the local file path, was recorded in structured JSON format for downstream processing.

The 60:20:20 train/validation/test split was applied to the augmented dataset as follows. Starting from the original single image per title class, ten augmented variants were generated, yielding a total of 11 images per title (1 original + 10 augmented). The split was then performed at the original-image level, not the augmented-variant level: for each title class, the original image was deterministically assigned to the training set, and the ten augmented variants were also retained exclusively in the training set. The validation and test sets were constructed from entirely separate original images obtained for 20% and 20% of classes respectively through independent data acquisition queries. This design ensures no augmented variant of any training image appears in the validation or test sets, preventing data leakage across splits. All five independent training runs used the same fixed train/validation/test partition, varying only the random seed for weight initialization and augmentation sampling order during training.

Augmentation Pipeline Design: We applied four groups of data augmentations to simulate museum environments, device artifacts, user-induced errors, and regularization effects.

- **Museum Environment Simulation:** To replicate the visual complexity of museum settings, we applied augmentations that simulate uneven illumination and local contrast changes, including secondary effects such as glass reflections and visitor-cast shadows.
- **Device-Induced Distortions:** To simulate artifacts commonly introduced by mobile devices, we incorporated distortions that mimic wide-angle lens effects (barrel and pincushion deformation), projective transformations, resolution degradation, and sensor noise.
- **User Error Robustness:** To enhance robustness against user-induced capture errors, we simulated motion blur, defocus artifacts, slight rotations, and partial crops.
- **Regularization *via* Region Dropout:** To prevent over-reliance on specific visual regions, we applied spatial dropout by randomly masking image portions, encouraging distributed feature representations.

Figure 1 illustrates the complete framework workflow from smartphone input image to final artwork prediction, showing the augmentation pipeline, Stage 1 department classifier, department assignment, Stage 2 department-specific title classifiers, and final Top-K ranked output.

Model Architecture

Our backbone is InceptionResNetV2 [16], whose multi-scale architecture captures diverse visual features across spatial and semantic levels. A lightweight Style-Aware Feature Fusion (SAFF) module is mounted atop the backbone, reweighting feature channels based on a learned 256-dimensional style embedding.

Backbone: InceptionResNetV2: We adopted InceptionResNetV2 as our backbone feature extractor. This architecture combines the multi-scale convolutional design of the Inception module [17] with the residual connections of ResNet [18]. Pre-trained on ImageNet, InceptionResNetV2 produces high-dimensional feature maps containing both low-level and high-level visual information.

Style-Aware Feature Fusion (SAFF) Module: To better capture artist-specific stylistic characteristics, we introduced the SAFF module, which operates in three steps.

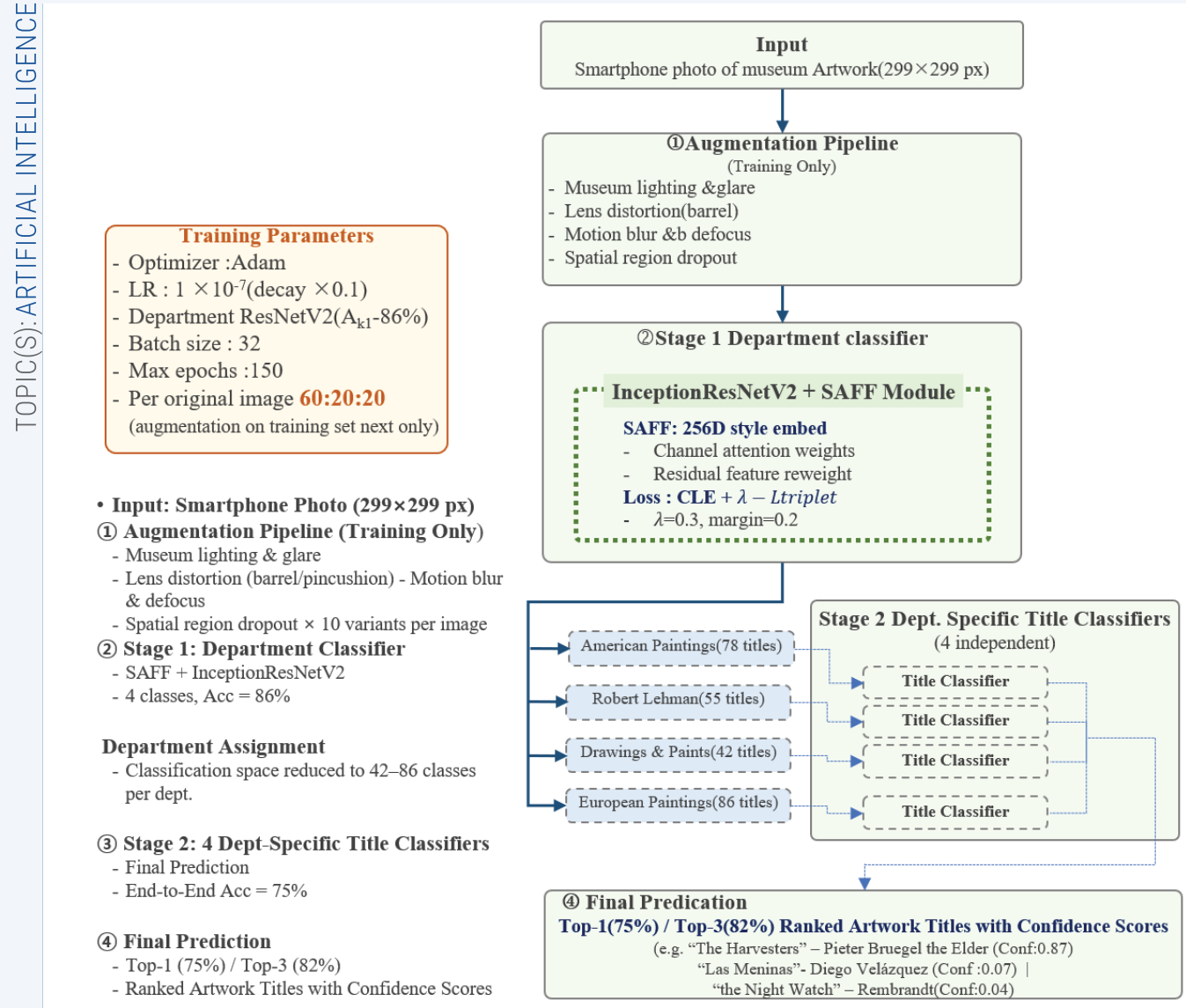

Figure 1 Complete framework workflow from input smartphone photograph to final artwork prediction. Augmentation (×10 variants) is applied only to the training set. Stage 1 assigns each image to one of four departments (reducing the classification space to 42-86 classes); Stage 2 applies a department-specific title classifier to produce the final ranked prediction.

- **Step 1** — Style Embedding Extraction: Global average pooling is applied to the InceptionResNetV2 feature map, and the result is passed through a fully connected layer to obtain a 256-dimensional style vector s.

- **Step 2** — Attention Weight Generation: The style vector is processed by a sigmoid-activated dense layer to produce channel-wise attention coefficients a = σ(Ws · s + b).

- **Step 3** — Residual Feature Fusion: The original feature map F is reweighted: F' = F + F ⊙ Reshape(a), where ⊙ denotes element-wise multiplication.

The model is trained end-to-end using a combined objective: L = LCE + λ · Ltriplet, where λ = 0.3 and the triplet loss margin is 0.2.

Two-Stage hierarchical classification: Stage 1 — Department Classification: An SAFF-augmented InceptionResNetV2 backbone classifies each artwork into one of four departments: European Paintings (86 titles), American Paintings and Sculpture (78 titles),

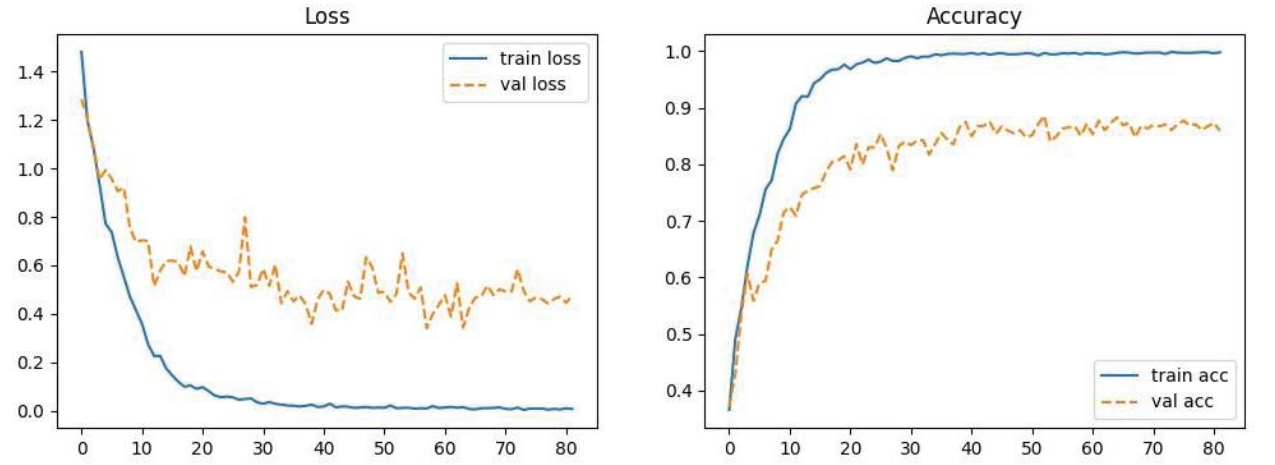

Figure 2 Training and validation loss and accuracy curves for the department-level classifier.

Robert Lehman Collection (55 titles), and Drawings and Prints (42 titles).

Stage 2 — Department-Specific Title Recognition: Four independent title classifiers are trained, each fine-tuned on the subset of images belonging to its respective department, reducing overfitting under the few-shot constraint.

Both Stage 1 and Stage 2 models were trained using the Adam optimizer with an initial learning rate of 1×10^{-3}, reduced by a factor of 0.1 at epochs 40 and 80. The batch size was set to 32 for Stage 1 training and 16 for Stage 2 (due to smaller per-department sample counts). Training ran for a maximum of 150 epochs with early stopping on validation loss (patience = 20 epochs); Stage 1 converged at epoch 82 and Stage 2 models converged between epochs 45 and 91 depending on department. Stage 1 was trained first to convergence before Stage 2 fine-tuning commenced, following a sequential training protocol. The combined loss weight $\lambda = 0.3$ and triplet margin = 0.2 were selected *via* grid search on the validation split.

Results and Discussion

Dataset and class counts

The original dataset comprises eight departments; four departments with very few title classes (Asian Art: 12, Modern and Contemporary Art: 12, Musical Instruments: 1, and Islamic Art: 1) were excluded to ensure adequate per-class sample sizes and training stability. The remaining four departments are shown in table 1.

Department-level classification

As shown in table 2, the proposed model achieves a test accuracy of 0.86 (86%) and a weighted F1-score of 0.86, with a final triplet loss of 0.62. All results in tables 2-4 are averaged over five independent training runs;

Table 1: Title class distribution across the four selected departments.

Department	Number of Title Classes
European Paintings	86
American Paintings and Sculpture	78
Robert Lehman Collection	55
Drawings and Prints	42
Total	**261**

Table 2: Department-level classification performance (4 classes).

Model	Test Acc	Test Loss	F1-Score	Params (M)
Proposed Model (InceptionResNetV2 + SAFF)	0.86	0.62	0.86	59.6

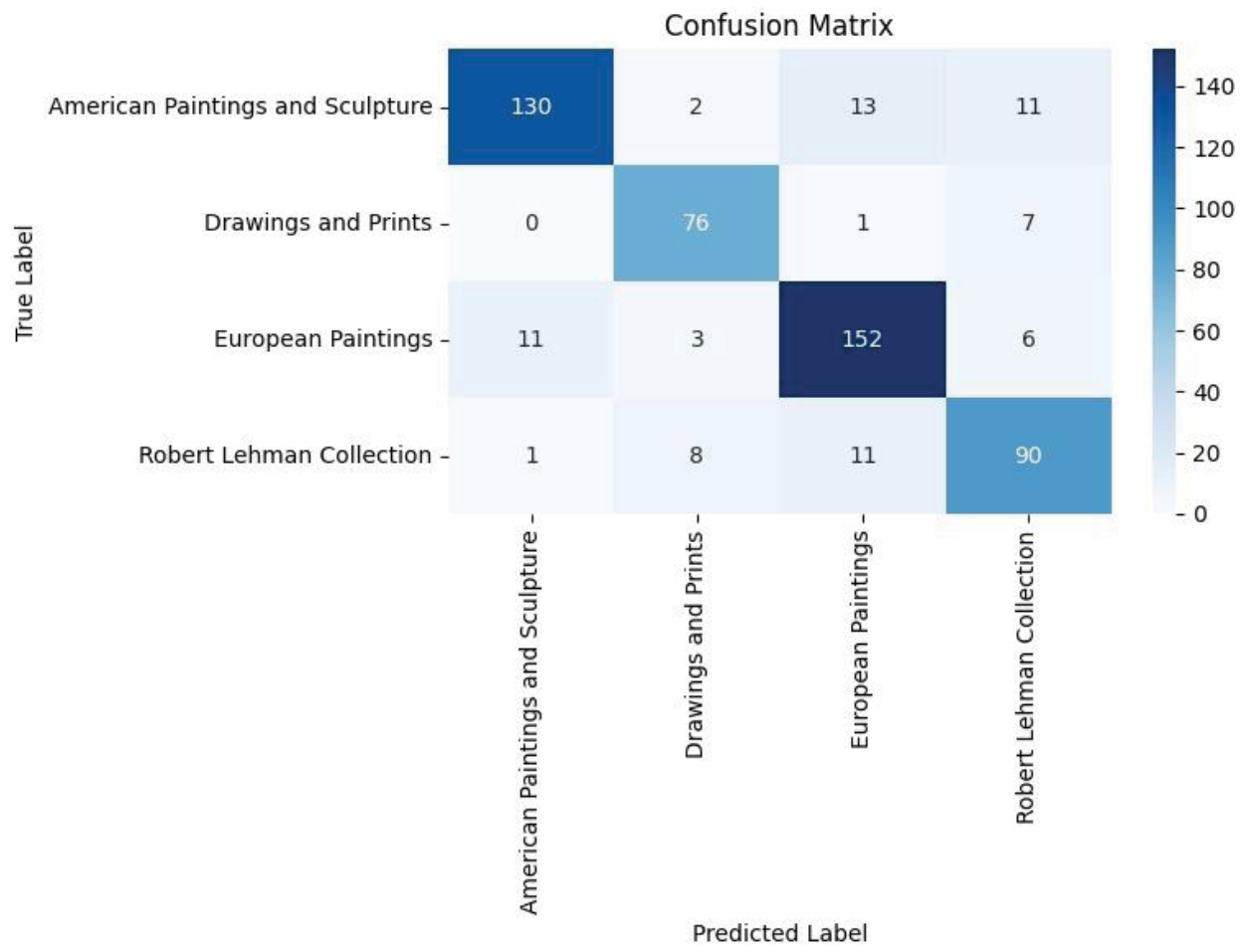

Figure 3 Confusion matrix for department-level classification.

the department-level test accuracy corresponds to a mean of 0.857 ± 0.012, and the end-to-end Top-1 accuracy of 0.75 corresponds to a mean of 0.748 ± 0.018 (Figure 2).

Figure 3 presents the confusion matrix for department-level classification. European Paintings and American Paintings and Sculpture show high correct prediction counts. Misclassifications are more frequent in the Robert Lehman Collection, possibly due to its diverse stylistic influences.

Figure 4 shows t-SNE visualization of style vector representations, exhibiting well-separated clusters aligned with department labels, confirming that the style-aware architecture facilitates robust feature disentanglement.

Title-level classification

Table 3 reports the test accuracies across the four departments. The proposed model consistently achieves ≥0.70 accuracy within each department, with a weighted average of 0.84 across all 261 title classes and a Top-3 accuracy of 0.91 (Figure 5).

End-to-end classification vs. baseline models

Table 4 shows end-to-end classification results. The proposed model achieves an overall Top-1 accuracy of 0.75 (75%) across all 261 title classes, substantially outperforming flat baselines. The ResNet50 baseline collapsed to 0.00 accuracy — a pattern consistent with output-layer saturation under a flat 261-way Softmax combined with extremely small per-

Table 3: Title-level classification performance per department.

Model	Department	Test Acc	Test Loss	F1 (%)	Params (M)	Top-3 Acc
Proposed (InceptionResNetV2+SAFF)	European Paintings	0.85	0.87	0.84	59.6	0.91
	American Paintings & Sculpture	0.91	0.54	0.91	59.2	0.94
	Robert Lehman Collection	0.70	1.28	0.59	59.6	0.88
	Drawings and Prints	0.86	0.70	0.85	59.6	0.90
	Weighted Average	0.84	0.83	0.81	59.5	0.91
Backbone Only (InceptionResNetV2)	European Paintings	0.42	3.39	0.40	58.1	0.56
	American Paintings & Sculpture	0.59	2.24	0.57	58.0	0.72
	Robert Lehman Collection	0.50	2.06	0.49	56.4	0.67
	Drawings and Prints	0.58	1.96	0.54	58.0	0.71
	Weighted Average	0.55	2.00	0.51	58.0	0.72

Table 4: End-to-end classification performance across all 261 title classes.

Model	Test Acc	Test Loss	F1-Score	Top-3 Acc
Proposed (InceptionResNetV2 + SAFF + Hierarchical)	**0.75**	**0.65**	**0.78**	**0.82**
InceptionResNetV2 (flat)	0.56	2.26	0.54	0.68
InceptionV3 (flat)	0.45	2.76	0.41	0.60
ResNet50 (flat)	0.00	5.63	0.00	0.02
VGG16 (flat)	0.37	3.14	0.35	0.50

Table 5: Ablation study of SAFF components (department-level, 4-class classification).

Configuration	Test Acc	F1-Score	Δ
Full SAFF (proposed)	0.86	0.86	—
w/o Style Embedding (+ w/o Triplet Loss)	0.77	0.76	−0.09
w/o Channel Attention Weights	0.79	0.78	−0.07
w/o Residual Fusion (hard replace)	0.80	0.79	−0.06
w/o Triplet Loss (only)	0.83	0.82	−0.03
Backbone Only (no SAFF)	0.72	0.70	−0.14

class sample counts — reinforcing the necessity of hierarchical decomposition.

Ablation study and attention mechanism comparison

Table 5 reports department-level Top-1 accuracy when each SAFF component is removed. Removing the style embedding causes the largest accuracy drop (−0.09). Table 6 compares SAFF against established attention mechanisms; SAFF achieves the highest accuracy (0.86) and F1-score (0.86).

Few-shot learning evaluation

Table 7 compares our framework against established few-shot and metric-learning methods adapted to the same dataset. Our proposed hierarchical model substantially outperforms Relation Networks (0.63), Matching Networks (0.58), and Prototypical Networks (0.61), achieving a Top-1 accuracy of 0.75.

Table 6: Comparison of attention mechanisms (department-level, 4-class classification).

Attention Mechanism	Test Acc	F1-Score	Extra Params (M)
SAFF (proposed)	0.86	0.86	+1.5
SE-Net	0.82	0.81	+1.2
CBAM	0.83	0.82	+1.8
Self-Attention (Transformer)	0.81	0.80	+3.4

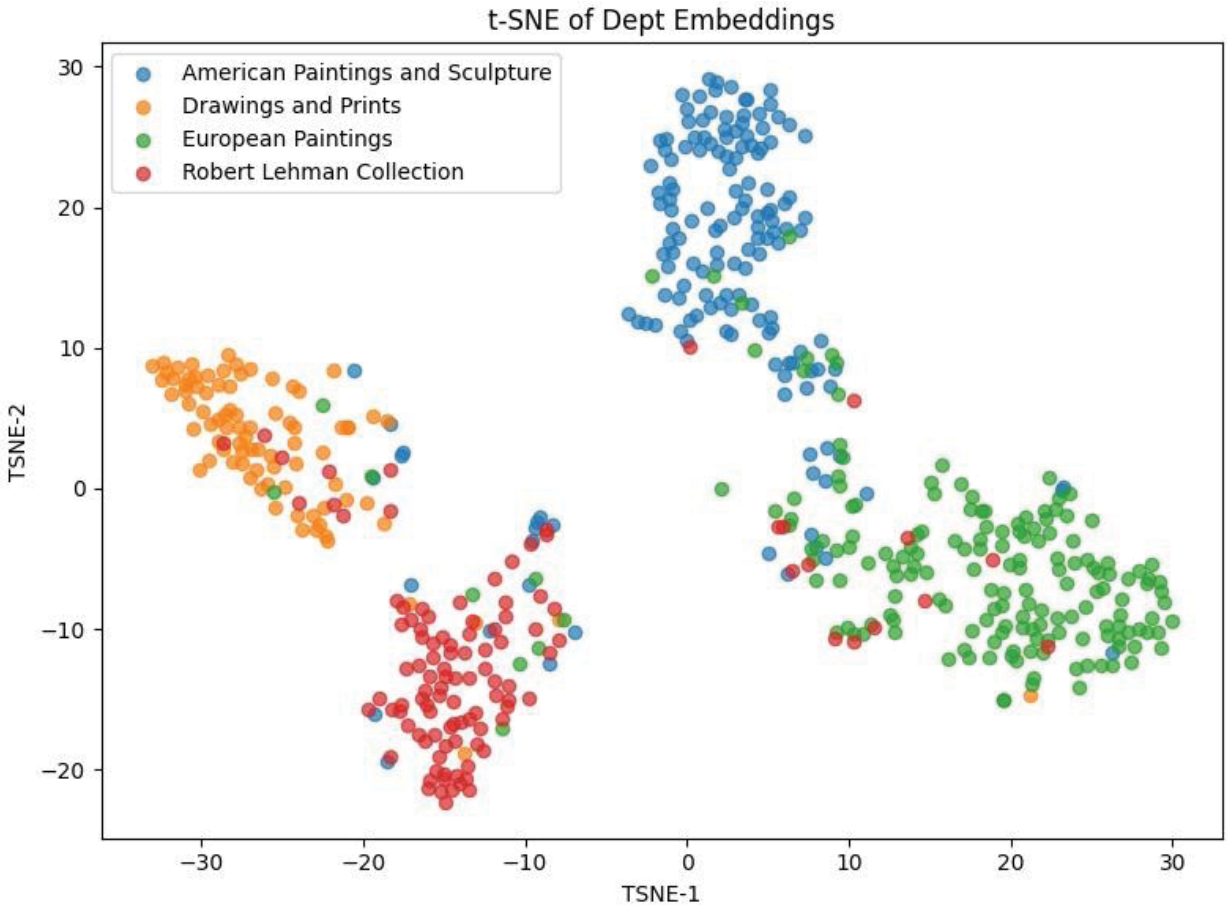

Figure 4 t-SNE visualization of style vectors for department-level classification. Each color corresponds to one department.

Generalization *vs.* augmentation dependency: The department-level accuracy on the standard test split is 0.86, while accuracy on the unaugmented holdout set is 0.81 - a gap of 0.05, suggesting the model does not overfit to augmentation-specific artifacts.

In comparison with recent studies, our framework outperforms the Prototypical Network baseline [12] by 12 percentage points (0.75 *vs.* 0.63) and the ViT-B/16 fine-tuning approach of Condoleo, et al. [13] (which achieved 71% under full supervision with far more training images per class) by 4 percentage points under a much more restrictive one-shot-per-title constraint. Yanagi, et al. [14] style-augmented prototypes achieved improvements of up to 8 pp over standard prototypical networks — consistent with our ablation result showing that removing the style embedding alone costs 9 pp (0.86 → 0.77). Together, these comparisons confirm that our framework advances the state of the art specifically at the intersection of style-aware feature learning, museum-domain

Figure 5 Top-4 nearest neighbors by Euclidean distance in the learned style embedding space for a representative query image.

Table 7: Comparison with few-shot and metric-learning approaches (end-to-end Top-1 accuracy, 261 classes).

Method	Top-1 Acc	Top-3 Acc	F1-Score
Proposed (InceptionResNetV2 + SAFF + Hierarchical)	**0.75**	**0.82**	**0.78**
Relation Networks [adapted]	0.63	0.74	0.61
Prototypical Networks [adapted]	0.61	0.72	0.59
Matching Networks [adapted]	0.58	0.70	0.55

Table 8: Computational efficiency comparison (299×299 input).

Model	Inference (ms)	FLOPs (G)	Memory (MB)	End-to-End Acc
InceptionResNetV2 + SAFF (proposed)	38	13.1	612	0.75
MobileNetV3-Large + SAFF	12	0.45	148	0.61
EfficientNet-B0 + SAFF	18	0.39	192	0.64
InceptionResNetV2 (flat baseline)	35	12.8	598	0.56

augmentation, and hierarchical classification under extreme data scarcity.

Computational efficiency

The proposed model requires 38 ms per image on GPU, suitable for museum kiosk applications. Lightweight alternatives reduce inference time substantially, with INT8 quantization yielding approximately 8 ms CPU inference (Table 8).

Limitations

Several limitations qualify the scope of these findings. First, all experiments are conducted on a single institutional collection (the Metropolitan Museum of Art); generalization to museums with different cataloguing conventions, imaging equipment, or artwork media has not been verified. Second, the synthetic augmentations have not been validated against a held-out set of genuine smartphone photographs taken inside a museum; a residual domain gap is possible and is not quantified in this study. Third, the 256-dimensional style embedding size and triplet-loss weighting coefficient ($\lambda = 0.3$) were selected *via* a single grid search. Finally, the Robert Lehman Collection consistently exhibits the weakest title-level performance (Test Acc. = 0.70), attributed to the heterogeneous stylistic composition of a privately assembled collection.

An important avenue for future validation is to evaluate the proposed framework using genuine smartphone photographs taken inside real museum environments. Such a real-world validation dataset could be constructed by collecting visitor-style photographs (varying angles, distances, lighting conditions, and partial occlusions) of artworks in the Metropolitan Museum of Art or comparable

institutions, then testing the trained models directly on these unconstrained images without any additional fine-tuning. This would directly quantify the residual domain gap between the synthetic augmentation distribution used in training and the real distribution of museum visitor photographs — the primary open question that the current study cannot answer with synthetic-only evaluation. Preliminary steps in this direction could include crowd-sourcing photographs *via* a dedicated museum application or partnering with museum accessibility programs that already collect visitor interaction data. Such external validation would be a prerequisite before deploying the system as a production accessibility tool.

Conclusion

We presented a style-aware hierarchical recognition framework for museum artwork that directly addresses the challenge of having only a single training image per title. The framework combines three mutually reinforcing components: (1) a domain-specific augmentation pipeline generating ten realistic, museum-aware synthetic variants per image; (2) a lightweight SAFF module that adaptively emphasizes stylistically discriminative channels; and (3) a two-stage hierarchical classifier predicting department then title. The proposed system achieves 86% department-level accuracy, 84% weighted-average Top-1 accuracy across 261 title classes, and 75% end-to-end Top-1 accuracy, substantially outperforming all flat baseline models.

Future work will pursue five directions: (1) real-time museum guidance system deployment *via* MobileNetV3-Large + SAFF integrated into a smartphone prototype; (2) mobile and edge device optimization *via* structured pruning and INT8 quantization-aware training; (3) cross-museum generalization extending to additional collections (Rijksmuseum, Louvre); (4) multimodal and metadata fusion incorporating artist biographical data, period labels, and descriptive text; and (5) personalized museum guidance integrating the recognition pipeline with a large language model for visitor-tailored artwork descriptions.

Acknowledgements

The authors thank the Metropolitan Museum of Art for providing open access to its collection data and imagery through the public BigQuery interface.

Funding Statement

The author(s) received no specific funding for this study.

Author Contributions

Conceptualization, Min Kim and Youngbok Cho; methodology, Min Kim; software, Min Kim; validation, Min Kim and Youngbok Cho; formal analysis, Min Kim; investigation, Min Kim; data curation, Min Kim; writing — original draft preparation, Min Kim; writing — review and editing, Youngbok Cho; supervision, Youngbok Cho; project administration, Youngbok Cho. All authors reviewed and approved the final version of the manuscript.

Data Availability Statement

The artwork images and metadata used in this study are publicly available *via* the Metropolitan Museum of Art Open Access initiative at https://www.metmuseum.org/about-the-met/policies-and-documents/open-access and the Met BigQuery public dataset at https://bigquery.cloud.google.com/dataset/metmuseum:museum.

Ethics Approval

Not applicable.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS 2012). Lake Tahoe (NV): NIPS; 2012. p. 1097-1105.
2. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas (NV): IEEE; 2016. p. 770-778. doi:10.1109/CVPR.2016.90.
3. Balakrishnan N, Lam SH, Yang P. Using convolutional neural networks to classify and understand artists from the Rijksmuseum. Stanford (CA): Stanford University; 2017. CS231n Course Report.
4. Artwork classification and recognition system based on convolutional neural network. Research report. 2022.
5. Wang Z, Song H. A fusion model for artwork identification based on convolutional neural networks and transformers [Preprint]. arXiv. 2025;2502.18083.
6. Senjam SS. Smartphones as assistive technology for visual impairment. Eye (Lond). 2021;35(8):2078-2080. doi:10.1038/s41433-021-01499-w.
7. Sharma P, Kaur R. A survey on smartphone-based assistive technologies for visually impaired people. AIP Conf Proc. 2023;2705(1):040002. doi:10.1063/5.0133329.
8. Lee A. From Braille to Be My Eyes—there's a revolution happening in tech for the blind [Internet]. London: The Guardian; 2017 Jun 26.
9. Sparks H. Houston Museum of Natural Science introduces app for people with low vision [Internet]. Axios. 2025 Feb 11.
10. Using convolutional neural networks to classify art genre [Honours thesis]. Townsville (AU): James Cook University; 2022.
11. Johnson K. AI could change how blind people see the world [Internet]. WIRED. 2023 Jul 5.
12. Elkholy M, El-Bendary N, Hassanien AE. Few-shot artwork classification using Prototypical Networks on the WikiArt dataset. Multimed Tools Appl. 2024;83(5):14231-14250. doi:10.1007/s11042-023-16213-5.
13. Condoleo F, Lombardi F, Ramasso E. Fine-grained artwork recognition with Vision Transformers under limited supervision. J Cult Herit. 2024;68:112-123. doi:10.1016/j.culher.2024.01.005.
14. Yanagi T, Hayashi K, Takimoto H. Style-augmented prototype learning for few-shot painting classification. Pattern Recognit Lett. 2025;180:45-52. doi:10.1016/j.patrec.2024.12.008.
15. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Albumentations: fast and flexible image augmentations. Information (Basel). 2020;11(2):125. doi:10.3390/info11020125.
16. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: Proc AAAI Conf Artif Intell. San Francisco (CA): AAAI Press; 2017. p. 4278-4284.
17. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proc IEEE Conf Comput Vis Pattern Recognit. Boston (MA): IEEE; 2015. p. 1-9. doi:10.1109/CVPR.2015.7298594.
18. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In: Lect Notes Comput Sci. Cham: Springer; 2016. p. 630-645. doi:10.1007/978-3-319-46493-0_38.