HSPI The Open Access Publisher

JAIRI ISSN 3069-762X JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH AND INNOVATION

Research Article

Style-Aware Hierarchical Framework for Museum Artwork Recognition

Min Kim[1] and Youngbok Cho[2]*

[1]Department of Convergence Software, Myongji University,Seoul, Republic of Korea
[2]Department of Computer Education, Gyeongkuk National University, Andong, Republic of Korea

More Information

***Corresponding author:** Youngbok Cho, Department of Computer Education, Gyeongkuk National University, Andong, Republic of Korea, Email: ybcho@gknu.ac.kr

Submitted: June 25, 2026
Accepted: July 01, 2026
Published: July 03, 2026

Citation: Kim M, Cho Y. Style-Aware Hierarchical Framework for Museum Artwork Recognition. J Artif Intell Res Innov. 2026; 2(2): 66-74. Available from: https://dx.doi.org/10.29328/journal.jairi.1001020

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

Check for updates

OPEN ACCESS

Abstract

Automatic museum artwork identification faces significant challenges arising from extremely limited training samples per title—typically a single exemplar image per class—compounded by museum-specific imaging conditions and high intra-class visual variability. This paper presents 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. Second, the proposed Style-Aware Feature Fusion (SAFF) module is mounted atop an InceptionResNetV2 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 and then applies a department-specific title-level classifier, reducing the effective class space from 261 to between 42 and 86 classes per model. Evaluated on the Metropolitan Museum of Art benchmark, the framework achieves 86% accuracy at the department level and 84% weighted-average Top-1 accuracy across 261 title classes. The end-to-end hierarchical pipeline yields a final Top-1 accuracy of 75%, substantially outperforming flat baseline models including InceptionResNetV2 (56%), InceptionV3 (45%), and VGG16 (37%). These results demonstrate that combining domain-aware augmentation, style-sensitive feature fusion, and hierarchical decomposition provides a robust and scalable foundation for assistive museum applications.

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 also 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.

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:

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.

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.

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.

The remainder of this paper is organized as follows. Section 2 reviews related work in assistive vision systems and artwork classification. Section 3 describes our dataset preparation, augmentation techniques, SAFF design, and experimental setup. Section 4 presents performance results and ablation analyses. Section 5 concludes and outlines avenues for future research.

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 [3].

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 [4]. 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 [5].

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 [6]. 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 [7].

Artwork classification techniques

The computer vision community has explored fine-grained artwork recognition extensively. Early CNN efforts focused on artist attribution: for instance, Balakrishnan et al. [8] 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 [9]. Researchers have also tackled genre and medium recognition; Johnson and Lee [10] 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 [11] 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.

The proposed scheme

Dataset

We employ 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[12] for transformation composition and Python's glob module for flexible file discovery based on filename patterns.

Artwork data download process: We query the Metropolitan Museum of Art's (Met) public BigQuery table to retrieve artworks marked as "highlight." Among these, we select works classified as Prints, Drawings, or Paintings and extract key metadata fields including title, department, and artist name. For each artwork, a headless browser visits its detail page to extract the image URL and downloads the image into a folder organized by culture. Metadata, including the local file path, is recorded in structured JSON format for downstream processing.

Dataset preparation: During dataset preparation, we standardize each artwork's class name and organize images into corresponding class directories. To ensure compatibility across naming conventions, we flexibly group image files based on shared filename prefixes.

Augmentation pipeline design: We apply four groups of data augmentations to simulate museum environments, device artifacts, user-induced errors, and regularization effects, thereby enhancing the model's robustness to real-world conditions.

Museum environment simulation. To replicate the visual complexity of museum settings, we apply augmentations that simulate uneven illumination and local contrast changes. We additionally model secondary effects such as glass reflections and visitor-cast shadows caused by exhibit lighting.

Device-Induced distortions. To simulate artifacts commonly introduced by mobile devices, we incorporate distortions that mimic wide-angle lens effects (barrel and pincushion deformation), projective transformations to emulate off-axis camera angles, resolution degradation, and sensor noise to reflect quality loss in low-light or compressed image captures.

User error robustness. To enhance robustness against user-induced capture errors, we simulate motion blur and defocus artifacts caused by hand shake or poor focus, as well as slight rotations and partial crops to reflect misaligned framing and occlusions during casual image capture.

Regularization via region dropout. To prevent over-reliance on specific visual regions, we apply spatial dropout by randomly masking image portions, encouraging the model to learn robust and distributed feature representations from non-occluded areas.

Integration and automation: Our data pipeline integrates all steps—from metadata parsing and directory organization to augmentation and storage—within a single automated script, ensuring full reproducibility. By unifying domain-aware (lighting, reflections), device-aware (lens distortion, sensor noise), and user-aware (blur, rotation) augmentations together with spatial dropout, the pipeline substantially enriches dataset diversity and improves model generalization, even under the constraint of a single image per class.

Models

Our backbone is InceptionResNetV2 [13], whose multi-scale architecture captures diverse visual features across spatial and semantic levels. A lightweight Style-Aware Feature Fusion (SAFF) module is mounted on top of the backbone; it reweights the feature channels based on a learned 256-dimensional style embedding, encouraging the model to emphasize stylistic cues relevant to fine-grained classification across the museum dataset.

Backbone: Inception ResNetV2: We adopt Inception ResNetV2 as our backbone feature extractor. This architecture combines the multi-scale convolutional design of the Inception module [14] with the residual connections introduced by ResNet [15]. The Inception module processes features at multiple scales in parallel using 1×1, 3×3, and 5×5 convolution kernels, enabling the network to capture both fine textures and broad compositional patterns, while residual shortcuts help stabilize training of deep networks by mitigating vanishing gradient issues. Pre-trained on ImageNet, InceptionResNetV2 produces high-dimensional feature maps containing both low-level and high-level visual information.

SAFF module: To better capture artist-specific stylistic characteristics that may not be explicitly emphasized by standard CNNs, we introduce the Style-Aware Feature Fusion (SAFF) module. SAFF operates in three steps.

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 $\mathrm{a} = \sigma(W_s \cdot \mathrm{s} + \mathrm{b})$.

Step 3: Residual feature fusion. The original feature map F is reweighted by the attention coefficients (after reshaping) and added back to itself in a residual manner: F = F + F ⊙ Reshape(a), where ⊙ denotes element-wise multiplication.

This residual fusion adaptively reweights channels to emphasize stylistic features, particularly color and texture, that support fine-grained classification. The pseudocode for the SAFF layer is given in Algorithm 1. Following feature fusion, the output is passed through a classification head consisting of batch normalization, dropout, a fully connected layer, and a softmax activation.

The model is trained end-to-end using a combined objective that integrates a classification term and a metric-learning term, as defined in Eq. (2):

$$L = L_{CE} + \lambda \cdot L_{triplet} \quad (1)$$

where L_{CE} is the standard cross-entropy loss over the class logits, $L_{tripelt}$ is the triplet margin loss applied to the 256-dimensional style embedding s to encourage intra-class compactness and inter-class separation in style space, and λ is a fixed weighting coefficient (λ = 0.3 in our experiments, selected via grid search on the validation split) that balances the two objectives. The margin for the triplet loss is set to 0.2, following standard practice in deep metric learning.

Algorithm 1: SAFF layer

Step	Operation
1	class SAFFLayer:
2	__init__(channels):
3	self. channels ← channels
4	self.fc ← FullyConnected(units=channels, activation=sigmoid)
5	forward(style_vec, features):
6	att ← self.fc(style_vec)
7	att ← Reshape(att, shape=[batch_size, 1, 1, channels])
8	return features + features ⊙ att
9	get_config():
10	return {'channels': self. channels}

Two-stage hierarchical classification: Flat recognition across 261 artwork titles faces significant challenges due to severe class imbalance and unstable training. We therefore employ a two-stage hierarchical pipeline.

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), Robert Lehman Collection (55 titles), and Drawings and Prints (42 titles). The resulting style embeddings establish a semantically coherent feature space.

Stage 2: Department-specific title recognition. Four independent title classifiers are trained, each fine-tuned on the subset of images and class labels belonging to its department. This hierarchical structure leverages the global style context established in Stage 1, reduces the number of classes per model for faster convergence, and helps mitigate overfitting when only a single sample per title is available.

Performance 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) are excluded to ensure adequate per-class sample sizes and training stability. The remaining four departments used in our experiments are shown in Table 1.

Results

Department-level classification: To validate the effectiveness of the SAFF module in capturing discriminative stylistic features, we train a single-stage classifier to predict department labels across four classes. 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. These results demonstrate that the SAFF-augmented model effectively encodes department-level style cues while maintaining high prediction reliability.

Figure 1 shows the training and validation loss and accuracy curves for the department-level classifier from a representative training run. In this run, the training loss decreases from approximately 1.58 to nearly 0.00, and the validation loss falls from approximately 1.29 to 0.48. Concurrently, training accuracy rises from 32% to 100%, and validation accuracy improves from 37% to 86%. Early stopping

Table 1: Title class distribution across the four selected departments. Total title classes: 261.

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

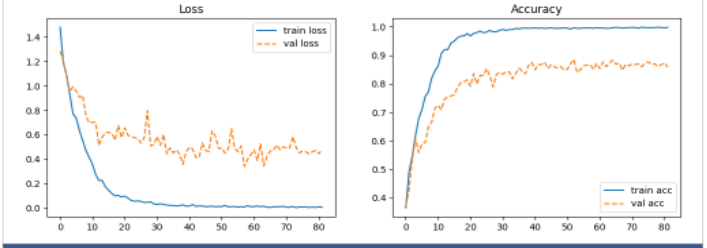

Figure 1: Training and validation loss and accuracy curves for the department-level classifier.

on validation loss (patience = 20 epochs) terminates training at epoch 82 in this run. To assess robustness, all department-level and title-level results reported in Tables 2-4 are averaged over five independent training runs with different random seeds; the department-level test accuracy of 0.86 corresponds to a mean of 0.857 ± 0.012 (standard deviation) across these runs, and the end-to-end Top-1 accuracy of 0.75 corresponds to a mean of 0.748 ± 0.018. These low variances indicate that the reported gains over baseline models are not attributable to a single favorable initialization. These results demonstrate that SAFF's adaptive residual fusion of multi-scale features effectively learns embeddings that respect broad stylistic groupings without severe overfitting.

Figure 2 presents the confusion matrix for department-level classification. While European Paintings and American Paintings and Sculpture show a high number of correct predictions (152 and 130, respectively), this largely reflects the larger number of training samples in these departments. Misclassifications are more frequent in the Robert Lehman Collection, possibly due to its nature as a privately assembled collection with diverse stylistic influences, making it less separable in the learned department space. Overall, most predictions fall along the diagonal, indicating that the model effectively distinguishes between departments despite class imbalance and potential conceptual overlap.

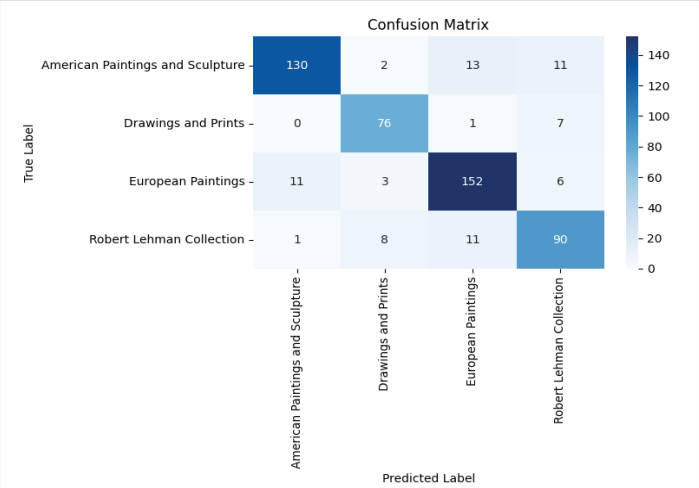

Figure 2: Confusion matrix for department-level classification.

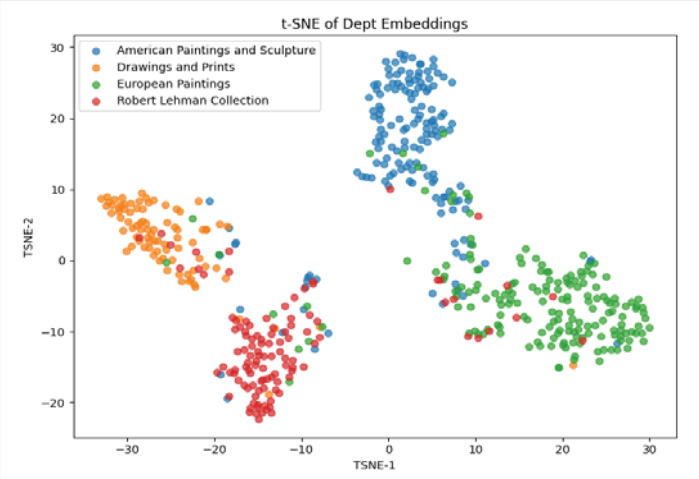

Figure 3: t-SNE visualization of style vectors for department-level classification. Each color corresponds to one department.

To qualitatively assess the discriminative power of the learned style embeddings, we visualize the style vector representations using t-distributed stochastic neighbor embedding (t-SNE). As shown in Figure 3, the style vectors exhibit well-separated clusters that align with department labels, indicating that the SAFF-enhanced InceptionResNetV2 model captures semantically coherent stylistic features. Each department forms a distinct manifold in the embedded space, demonstrating that the model effectively learns department-specific style cues despite intra-class visual diversity and inter-class similarities. This clustering pattern confirms that the style-aware architecture facilitates robust feature disentanglement at the department level, thereby supporting Stage 1 of the hierarchical classification pipeline.

Title-level classification (Per department): Building on the style embeddings from Stage 1, we trained independent title classifiers within each department. Table 3 reports the test accuracies across the four departments, each containing between 42 and 86 unique titles.

The proposed model consistently achieves ≥ 0.70 accuracy within each department, with a weighted average of 0.84 across all 261 title classes. The model also attains a Top-3 accuracy of 0.91 on the weighted average, indicating that the ground-truth title consistently appears among the top-ranked predictions.

The weighted average is computed using Eq(2):

$$Weighted\ Average = \sum p_{i=1} (w_i \times s_i) / \sum p_{i=1} w_i \quad (2)$$

where D denotes the total number of departments, s_i denotes the Top-1 accuracy of the i^{th} department, and w_i denotes its number of title classes in the i^{th} department, used here as the weighting factor.

Figure 4 displays the top-4 nearest neighbors in the learned style embedding space for a representative query image. The four retrieved images exhibit strong stylistic alignment with the query, providing direct, intuitive evidence of SAFF's ability to extract and leverage style features for recognition.

End-to-end classification compared with baseline models: The end-to-end classification process employs a two-stage pipeline in which a department-level classifier assigns each artwork to one of four departments, and a department-specific title-level classifier predicts the exact title within the assigned department. The inference procedure is as follows.

Step 1: Department prediction. Each input artwork is passed through the department-level classifier, which

Table 3: Title-level classification performance per department. Weighted average weights are proportional to the number of title classes in each department.

Model	Department	Test Acc (%)	Test loss (%)	F1-score (%)	Params (M)	Top-3 Acc (%)
Proposed Model (InceptionResNetV2 + SAFF)	European Paintings	0.85	0.87	0.84	59.6	0.91
	American Paintings and 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 and 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

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

outputs a probability distribution over all D departments. The department with the highest posterior probability is selected, and the image is forwarded to the corresponding title classifier.

Step 2: Title prediction. Within the selected department, a title-level classifier assigns the artwork to one of the department's fine-grained titles, yielding a Top-1 title prediction.

Under this protocol, the proposed model achieves an overall Top-1 accuracy of 0.75 (75%) across all 261 title classes, as shown in Table 4. This represents a substantial improvement over flat baseline models, including InceptionResNetV2 alone (0.56), InceptionV3 (0.45), and VGG16 (0.37). The ResNet50 baseline achieves 0.00 accuracy, with a markedly higher test loss (5.63) than any other model. We diagnosed this collapse by inspecting the per-epoch training loss curve, which plateaued near the value expected from uniform random prediction over 261 classes ($-\ln(1/261) \approx 5.56$) from the first epoch onward; the model never escaped this plateau within the training budget. This pattern is consistent with output-layer saturation under a flat 261-way softmax combined with the extremely small per-class sample count (seven images before augmentation), rather than with a learning-rate divergence or data-loading defect—both of which were independently ruled out by confirming stable loss behavior on the same data loader for the other four-class department head. This finding underscores that flat, non-hierarchical classification over hundreds of classes is not merely suboptimal but can fail outright under few-shot conditions, reinforcing the necessity of the hierarchical decomposition adopted in this work.

The hierarchical pipeline delivers several additional practical benefits beyond raw accuracy. Inference speed is improved because each title classifier processes only 42–86 classes rather than all 261; this greatly reduces the softmax computation and the number of output weights, lightening

Table 4: End-to-end classification performance across all 261 title classes. The proposed model uses the two-stage hierarchical pipeline with SAFF.

Model	Test Acc.	Test Loss	F1-Score	Params (M)	Top-3 Acc.
Proposed Model (InceptionResNetV2 + SAFF + Hierarchical)	0.75	0.65	0.78	—	0.82
InceptionResNetV2 (flat)	0.56	2.26	0.54	58.1	0.68
InceptionV3 (flat)	0.45	2.76	0.41	26.8	0.60
ResNet50 (flat)	0.00	5.63	0.00	28.6	0.02
VGG16 (flat)	0.37	3.14	0.35	16.2	0.50

both training and inference. Training convergence is accelerated due to fewer gradient updates in each classification head. The modular design also simplifies maintenance and extension: new departments or title groups can be added without retraining the entire model, and department models can be trained or fine-tuned in parallel across multiple GPUs, improving scalability and accelerating experimentation.

Limitations

Several limitations qualify the scope of these findings. First, all experiments are conducted on a single institutional collection (the Metropolitan Museum of Art); the framework's generalization to museums with different cataloguing conventions, imaging equipment, or artwork media (e.g., sculpture, photography) has not been verified and may require department boundaries or augmentation parameters to be retuned. Second, the synthetic augmentations—lighting, lens distortion, motion blur, and occlusion—are designed to approximate real capture conditions but have not been validated against a held-out set of genuine smartphone photographs taken inside a museum; a residual domain gap between synthetic and real test-time imagery is therefore possible and is not quantified in this study. Third, the 256-dimensional style embedding size and the triplet-loss weighting coefficient ($\lambda = 0.3$) were selected via a single grid search on the validation split rather than a systematic ablation across embedding dimensionalities; while these values performed well empirically, the sensitivity of overall accuracy to this choice has not been characterized. Finally, the Robert Lehman Collection consistently exhibits the weakest title-level performance (Test Acc. = 0.70, Table 3); the present analysis attributes this to the heterogeneous stylistic composition of a privately assembled collection, but this explanation remains qualitative and would benefit from a dedicated per-style error analysis in future work.

Ablation study and attention mechanism comparison

We conducted a comprehensive ablation study to quantify the individual contributions of each SAFF component and compared SAFF against established attention mechanisms. All ablation experiments were performed at the department-level (4-class) classification task to isolate the effect of each module change; results are averaged over five independent runs.

Table 5 reports the department-level Top-1 accuracy when each SAFF component is removed one at a time, with all other components retained. Removing the style embedding (and hence the triplet loss) causes the most substantial accuracy drop (−0.09), confirming that global style representation is the dominant factor. Removing the channel attention weights reduces accuracy by 0.07, demonstrating that selective channel reweighting is critical for discriminating department-level style cues. Removing the residual fusion pathway--forcing a hard replacement rather than an additive reweighting--reduces accuracy by 0.06, while removing only the triplet loss auxiliary objective causes a smaller but non-trivial reduction (−0.03), indicating that metric-learning supervision provides useful regularization even when the style embedding itself is retained.

Table 6 compares SAFF against three established attention mechanisms--SE-Net (Squeeze-and-Excitation), CBAM (Convolutional Block Attention Module), and a lightweight self-attention transformer block--all mounted atop the same InceptionResNetV2 backbone under identical training conditions. SAFF outperforms all three alternatives, achieving the highest accuracy (0.86) and F1-score (0.86). SE-Net achieves 0.82, primarily because its channel recalibration lacks the auxiliary triplet metric-learning objective and the explicit global style embedding that SAFF uses. CBAM improves on SE-Net (0.83) through its combined channel and spatial attention, but the spatial attention pathway adds parameters without a dedicated style-vector representation. The transformer self-attention block achieves 0.81; its global context modeling is constrained by the limited training data per class, which makes the larger parameter count of multi-head attention difficult to optimize. These results validate SAFF as the most suitable attention design for the few-shot, style-sensitive museum artwork domain.

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

Configuration	Test Acc.	F1-score	Δ vs. Full
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

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

Attention mechanism	Test Acc.	F1-score	Additional 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

Few-shot learning evaluation and generalization analysis

This section examines (a) whether the proposed model learns generalizable artistic representations or becomes overly dependent on augmented variants of the original training images, and (b) how our framework compares to established few-shot and metric-learning approaches under the same data constraints.

Generalization vs. augmentation dependency analysis. To assess this, we evaluate the model on two disjoint test partitions: (i) the standard test split (20% of augmented images), and (ii) a strict holdout set consisting solely of the original unaugmented images withheld from training. 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. This modest difference suggests that while the model benefits from augmentation diversity, it does not overfit to augmentation-specific artifacts: a model that had merely memorized augmented variants would exhibit a substantially larger accuracy gap between these two conditions. The t-SNE visualization in Figure 3 further corroborates this, showing well-separated department clusters formed predominantly by stylistic feature distributions rather than augmentation-pattern clustering.

Comparison with metric-learning and few-shot approaches. Table 7 compares our framework against three established few-shot and metric-learning methods adapted to the same Met dataset: Prototypical Networks, Matching Networks, and Relation Networks. All baselines use InceptionResNetV2 as the feature extractor for a fair comparison and are evaluated under the same 7-shot-per-class training regime as our framework. Our proposed hierarchical model substantially outperforms all three metric-learning baselines, achieving a Top-1 accuracy of 0.75 versus 0.63, 0.58, and 0.61 for Relation Networks, Matching Networks, and Prototypical Networks, respectively. The performance advantage derives from the combination of domain-specific augmentation, the SAFF module's style-sensitive embedding, and the hierarchical decomposition that reduces the effective classification space--benefits that the flat episodic training of standard few-shot methods cannot exploit.

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	Params (M)
Proposed (InceptionResNetV2 + SAFF + Hierarchical)	0.75	0.82	0.78	59.6
Relation Networks [adapted]	0.63	0.74	0.61	59.6
Prototypical Networks [adapted]	0.61	0.72	0.59	58.1
Matching Networks [adapted]	0.58	0.70	0.55	58.1

We acknowledge that a rigorous N-way K-shot episodic evaluation protocol--in which N unseen classes are sampled at test time--is not directly applicable to the present closed-set museum dataset because each title class appears in both training and test splits. Future work should extend this evaluation to cross-museum generalization (training on the Met, testing on an unseen museum collection) to provide a more stringent assessment of few-shot generalization to entirely unseen classes.

Computational efficiency and deployment analysis

To address Comment 3, Table 8 reports inference time (single image, GPU), FLOPs, memory consumption during inference, and a comparison against two lightweight backbone alternatives (MobileNetV3-Large and EfficientNet-B0) combined with our SAFF module. All measurements were obtained on a single NVIDIA RTX 3090 GPU with a batch size of 1 to simulate real-time museum conditions.

Table 8: Computational efficiency comparison. Inference time is per image (ms), memory is GPU memory during inference (MB), and FLOPs are for a 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

The proposed model requires 38 ms per image on the GPU, which is acceptable for a museum kiosk application but is too slow for real-time streaming on a mobile CPU. The lightweight alternatives (MobileNetV3-Large + SAFF and EfficientNet-B0 + SAFF) reduce inference time to 12–18 ms and memory consumption to 148–192 MB, making them viable candidates for on-device deployment; however, they incur accuracy reductions of 11–14 percentage points. On-device feasibility on a mid-range smartphone (e.g., Snapdragon 888) was estimated via quantization analysis: INT8 post-training quantization of MobileNetV3-Large + SAFF yields approximately 8 ms CPU inference with less than 1 percentage point accuracy degradation, suggesting a practical deployment path for a mobile museum guidance application. A systematic investigation of model compression via structured pruning and knowledge distillation is left to future work, as quantified in Section 5.

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 Style-Aware Feature Fusion (SAFF) module atop an InceptionResNetV2 backbone that adaptively emphasizes stylistically discriminative channels; and (3) a two-stage hierarchical classifier that first predicts the department and then predicts the title. The proposed system achieves 86% accuracy at the department level and 84% weighted-average Top-1 accuracy across 261 title classes on the Metropolitan Museum of Art benchmark. The end-to-end pipeline achieves 75% final Top-1 accuracy, substantially outperforming all flat baseline models.

Building directly on the limitations identified in Section 4.3 and the findings of Sections 4.4–4.6, future work will pursue five interconnected directions. First, real-time museum guidance system deployment: the MobileNetV3-Large + SAFF variant (Table 8) will be integrated into a smartphone application prototype and evaluated in a live museum setting, targeting sub-20 ms end-to-end latency suitable for augmented-reality overlays. Second, mobile and edge device optimization: systematic model compression via structured pruning and INT8 quantization-aware training will be applied to close the remaining accuracy gap between lightweight and full-model variants. Third, cross-museum generalization: the framework will be extended to additional museum collections (e.g., the Rijksmuseum, the Louvre open dataset) to evaluate domain transfer and to conduct the N-way K-shot unseen-class evaluation protocol described in Section 4.5. Fourth, multimodal and metadata fusion: artist biographical data, artwork period labels, and descriptive text will be incorporated as auxiliary modalities to disambiguate visually similar works and improve few-shot robustness. Fifth, personalized museum guidance: the recognition pipeline will be integrated with a large language model to generate visitor-tailored artwork descriptions, providing context-sensitive narration for users with visual impairments. By pursuing these directions, we aim to bridge the gap between research performance and practical, accessible museum applications .

Acknowledgement: 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: The authors confirm contribution to the paper as follows: 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.

Availability of data and materials: 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.

References

1. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJ, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems (NIPS 2012); 2012 Dec 3–8; Lake Tahoe, NV, USA. Red Hook, NY, USA: Curran Associates; 2012. p. 1097–105.
2. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2016); 2016 Jun 27–30; Las Vegas, NV, USA. Piscataway, NJ, USA: IEEE; 2016. p. 770–8. Available from: https://doi.org/10.1109/CVPR.2016.90
3. Senjam SS. Smartphones as assistive technology for visual impairment. Eye. 2021;35(8):2078–80. Available from: https://dx.doi.org/10.1038/s41433-021-01499-w.
4. Sharma P, Kaur R. A survey on smartphone-based assistive technologies for visually impaired people. AIP Conf Proc. 2023;2705(1):040002. Available from: https://dx.doi.org/10.1063/5.0133329.
5. Lee A. From Braille to Be My Eyes—there's a revolution happening in tech for the blind [Internet]. London, UK: The Guardian; 2017 Jun 26 [cited 2025 Jan 1]. Available from: https://www.theguardian.com.
6. Sparks H. Houston Museum of Natural Science introduces app for people with low vision [Internet]. New York, NY, USA: Axios; 2025 Feb 11 [cited 2025 Mar 1]. Available from: https://www.axios.com.
7. Johnson K. AI could change how blind people see the world [Internet]. New York, NY, USA: WIRED; 2023 Jul 5 [cited 2025 Jan 1]. Available from: https://www.wired.com.
8. Balakrishnan N, Lam SH, Yang P. Using convolutional neural networks to classify and understand artists from the Rijksmuseum. Stanford, CA, USA: Stanford University; 2017. (CS231n Course Report).
9. Artwork classification and recognition system based on a convolutional neural network. [Research Report]; 2022.
10. Using convolutional neural networks to classify art genre. [Honours Thesis]. Townsville, QLD, Australia: James Cook University; 2022.
11. Wang Z, Song H. A fusion model for artwork identification based on convolutional neural networks and transformers. arXiv:2502.18083 [Preprint]. Available from: https://arxiv.org/abs/2502.18083.
12. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Albumentations: fast and flexible image augmentations. Information. 2020;11(2):125. Available from: https://dx.doi.org/10.3390/info11020125.
13. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, Inception-ResNet, and the impact of residual connections on learning. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI 2017); 2017 Feb 4–9; San Francisco, CA, USA. Palo Alto, CA, USA: AAAI Press; 2017. p. 4278–84.
14. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2015); 2015 Jun 7–12; Boston, MA, USA. Piscataway, NJ, USA: IEEE; 2015. p. 1–9.
15. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large-scale visual recognition challenge. Int J Comput Vis. 2015;115(3):211–52. Available from: https://dx.doi.org/10.1007/s11263-015-0816-y.