Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 1, No. 2 (2025)

Article

Enhanced CNN-based Feature Extraction and Classification for Chinese Artwork Styles

Jiaying Li [1,*]

[1] Integrated Marketing Communications, Northwestern University, Chicago, IL, USA
* Correspondence: Jiaying Li, Integrated Marketing Communications, Northwestern University, Chicago, IL, USA

Abstract: This research focuses on optimizing convolutional neural network (CNN) architectures for extracting and classifying visual features in traditional Chinese paintings, which represent a distinctive artistic tradition characterized by brushstroke techniques, ink variations, and compositional nuances. The proposed hierarchical feature extraction framework integrates multi-scale fusion strategies with specialized modules for brushstroke, color, and compositional analysis. By systematically comparing ResNet, VGG, and EfficientNet backbones and combining them with layer-wise fine-tuning, the methodology achieves superior performance with limited training samples. Experimental validation on collections of traditional Chinese paintings demonstrates significant improvements in accuracy over strong CNN baselines, with the best configuration increasing overall accuracy from 82.1% to 93.2%. The framework provides practical solutions for museum digitization and auction-house cataloging.

Keywords: convolutional neural networks; artistic feature extraction; transfer learning; Chinese painting classification

1. Introduction

1.1. Research Background and Motivation

1.1.1. Importance of AI in Artwork Style Recognition

The digital transformation of cultural heritage institutions demands automated analytical capabilities to process extensive art collections. The digitization of museum and auction-house collections has increased demand for automated, scalable tools to index and analyze large volumes of high-resolution images of artwork. In this study, we focus on two-dimensional paintings, particularly master Chinese ink paintings whose stylistic differences are often expressed through subtle brush-and-ink variations (e.g., stroke pressure, ink dispersion, and calligraphic line quality). Western oil paintings are discussed as a contrasting painting tradition to highlight the distinctive visual characteristics of Chinese ink paintings, in which color layering, texture buildup, and compositional density provide complementary visual cues for CNN-based feature learning. The application of these architectures to artistic domains offers unique opportunities to enhance curatorial workflows and authentication processes. Museums worldwide manage collections exceeding hundreds of thousands of items, creating pressing needs for scalable classification solutions [1].

1.1.2. Challenges in Chinese Artwork Feature Extraction

Chinese traditional paintings exhibit distinctive aesthetic characteristics that differentiate them from Western artistic traditions. The unique properties of brush-and-

Received: 08 November 2025
Revised: 01 January 2026
Accepted: 13 January 2026
Published: 18 January 2026

ink techniques produce subtle textural variations encoding artistic intent through pressure modulation, stroke velocity, and ink concentration gradients.

Limited availability of digitized Chinese artwork datasets compounds the technical difficulties. The small sample sizes typical of specialized collections necessitate sophisticated transfer learning strategies and data augmentation methods capable of generalizing from limited training data [2].

1.2. Research Objectives and Contributions

1.2.1. Optimization Goals for CNN-based Feature Extraction

This study is limited to two-dimensional paintings (i.e., digitized images of paintings) and does not address three-dimensional artifacts such as ceramics, bronzes, or sculpture. This investigation pursues three primary optimization objectives in artistic feature extraction. The first objective is to design hierarchical feature map structures specifically adapted to capture brushstroke textures, compositional arrangements, and color harmonies characteristic of Chinese painting traditions. The second objective involves systematic evaluation of transfer learning efficacy across multiple pre-trained architectures to identify optimal backbone networks for artistic classification tasks.

1.2.2. Key Technical Contributions

The research introduces several novel technical components that advance the state of the art in artistic feature extraction. A multi-scale feature fusion strategy aggregates representations from multiple network depths, enabling simultaneous capture of fine brushstroke details and holistic compositional structures. The specialized brushstroke enhancement module employs directional gradient analysis to amplify stroke-specific visual signatures.

1.2.3. Application Scenarios

The developed framework addresses practical requirements across multiple cultural heritage contexts. Museum institutions can leverage the technology to automate the cataloging of newly acquired artworks. Auction houses benefit from rapid style verification, which supports attribution decisions. Digital cultural heritage initiatives leverage classification capabilities to create searchable databases, enabling scholars to explore collections through style-based queries.

2. Related Work

2.1. Deep Learning for Artwork Classification

2.1.1. CNN Architectures in Art Style Recognition

The application of convolutional neural networks to artistic style recognition has evolved through multiple architectural generations. Early investigations adapted standard classification networks, such as AlexNet and VGG, to artistic datasets, achieving moderate success via transfer learning from ImageNet pretraining. Recent work has also explored contrastive pre-training (e.g., CLIP-Art) to learn more discriminative representations for fine-grained art classification, thereby benefiting style recognition under limited labeled date [3].

Recent research has explored specialized architectural modifications tailored to artistic visual characteristics. Attention mechanisms have been incorporated to focus network processing on visually distinctive regions within artworks. Multi-branch architectures process images at varying resolutions, capturing both fine artistic details and overall compositional structures [4].

2.1.2. Feature Map Hierarchy Design

The hierarchical organization of convolutional feature maps plays a critical role in artistic representation learning. Lower network layers typically capture edge orientations, color contrasts, and basic textures relevant to brushstroke analysis. Middle layers encode

more complex patterns, including compositional elements and spatial relationships. Higher layers develop abstract style representations, enabling discrimination between artistic movements [5].

2.2. Transfer Learning in Visual Art Analysis

2.2.1. Comparison of ResNet, VGG, and EfficientNet

Transfer learning leverages knowledge acquired from large-scale datasets to accelerate training and improve performance on specialized target tasks. The choice of pre-trained backbone architecture significantly influences artistic classification outcomes. VGG networks employ uniform architectural patterns with small convolutional kernels, providing straightforward feature extraction pipelines. ResNet architectures introduce skip connections, enabling gradient flow through dense networks [6].

EfficientNet represents a more recent architectural family optimizing network depth, width, and resolution through principled scaling methods. Comparative studies on artistic datasets have revealed varying performance characteristics across these architectures. EfficientNet variants often achieve superior accuracy per parameter count, offering computational efficiency advantages [7].

2.2.2. Fine-tuning Strategies for the Art Domain

Effective transfer learning requires careful calibration of fine-tuning strategies to balance the retention of pre-trained knowledge with adaptation to target-domain characteristics. Layer freezing approaches maintain lower-layer parameters while adapting higher layers, preserving general visual representations while learning style-specific features.

Layer-wise fine-tuning with discriminative learning rates applies different update magnitudes to varying depths within the network. Lower layers receive smaller learning rates to preserve edge and texture detectors, while higher layers undergo larger updates to adapt abstract representations to artistic styles [8].

2.2.3. Cross-domain Feature Transfer

The transfer of visual features between natural and artistic domains poses unique challenges due to differences in the distributions of visual statistics. Natural photographs exhibit different color distributions, texture patterns, and compositional conventions compared to painted artworks. Despite these differences, pre-trained networks capture visual primitives that remain useful for artistic analysis.

2.3. Data Augmentation and Few-shot Learning

2.3.1. Traditional Augmentation Techniques

Standard data augmentation methods apply geometric transformations and color perturbations to generate synthetic training examples from limited source datasets. Rotation, translation, and scaling operations increase sample diversity while preserving artistic content. Color jittering modifies brightness, contrast, and saturation to simulate imaging variations without altering fundamental artistic characteristics.

2.3.2. GAN-based Data Generation

Generative adversarial networks offer sophisticated augmentation capabilities beyond simple geometric transformations. GANs trained on artistic datasets learn to synthesize novel artworks that exhibit statistical properties matching the training distributions. Beyond standard CNN baselines, prior studies have explored separating style-relevant cues via difference-component modeling to better distinguish painting styles, indicating that explicitly emphasizing discriminative components can improve classification robustness [9].

2.3.3. Meta-learning Approaches for Limited Samples

Meta-learning frameworks address few-shot classification challenges by learning to learn from minimal examples. Prototypical networks represent each style category through prototype embedding computed from available training examples. Brushstroke-level representations have been used to distinguish between Chinese and Western painting images, highlighting that stroke patterns provide strong style cues that can complement global CNN features [10].

3. Proposed Method

3.1. Overall Framework Architecture

3.1.1. Input Preprocessing Pipeline

The preprocessing pipeline transforms raw artwork images into standardized representations suitable for network processing. Images are normalized to 512×512 pixels, balancing computational efficiency with the preservation of fine visual details necessary for brushstroke analysis. Aspect-ratio preservation through padding helps maintain stroke geometry and local orientation patterns, which are essential for downstream brushstroke-sensitive analysis and extraction methods [11]. Color space considerations significantly influence the quality of feature extraction. RGB color spaces provide standard representations compatible with pre-trained network expectations. Additional color space transformations to HSV coordinates isolate hue, saturation, and value components. Normalization procedures align input statistics with pre-training data distributions.

3.1.2. Feature Extraction Network Design

The core feature-extraction network adopts a hierarchical architecture that processes inputs through a sequence of convolutional blocks with progressively increasing abstraction levels. The network backbone is based on pre-trained ResNet, VGG, or EfficientNet architectures, providing robust visual feature extraction.

The network produces feature maps at multiple spatial resolutions corresponding to different semantic levels (Table 1). Shallow layers generate high-resolution feature maps capturing fine brushstroke textures. Intermediate layers produce moderate-resolution representations encoding compositional elements. Deep layers produce low-resolution feature maps that capture global artistic style characteristics.

Table 1. Comparison of Backbone Network Architectures for Artistic Feature Extraction.

Architecture	Depth (Layers)	Parameters (M)	Pre-training Dataset	Relative Downsampling Stages	Computational Cost (GFLOPs)
VGG16	16	138.4	ImageNet-1K	512→256→128→64→32	15.5
ResNet50	50	25.6	ImageNet-1K	512→256→128→64→32	4.1
ResNet101	101	44.5	ImageNet-1K	512→256→128→64→32	7.8
EfficientNet-B3	82	12.2	ImageNet-1K	512→256→128→64→32	1.8
EfficientNet-B5	120	30.4	ImageNet-21K	512→256→128→64→32	9.9

GFLOPs are reported under the standard 224×224 input setting from commonly used model specifications; computation increases approximately quadratically when using 512 × 512 inputs in our experiments.

3.2. Hierarchical Feature Extraction Optimization

3.2.1. Multi-scale Feature Fusion Strategy

The multi-scale fusion module aggregates feature representations from multiple network depths to create comprehensive artistic descriptors. The fusion strategy employs spatial upsampling of lower-resolution feature maps to match the dimensions of higher-resolution counterparts, followed by channel-wise attention weighting to emphasize informative feature channels [12].

Given feature maps F_i from layer i with spatial dimensions $H_i \times W_i$ and C_i channels, the fusion process computes:

$F_{upsampled_i} = Upsample(F_i, size=(H_{target}, W_{target}))$

Where upsampling employs bilinear interpolation. The attention mechanism computes channel importance weights:

$alpha_{i,c} = sigmoid(w_{i,c} \cdot GlobalAvgPool(F_{upsampled, i,c}))$

The final fused representation combines weighted features:

$F_{fused} = sum_i (alpha_i \odot F_{upsampled_i})$

The architecture diagram (Figure 1) illustrates the complete processing pipeline from input artwork to style classification. The visualization employs a layered block diagram representation with distinct color coding for different functional modules. The input stage shows a Chinese painting image (512×512 resolution) entering the preprocessing pipeline, depicted as a series of transformation blocks including resolution normalization, color space conversion, and statistical normalization.

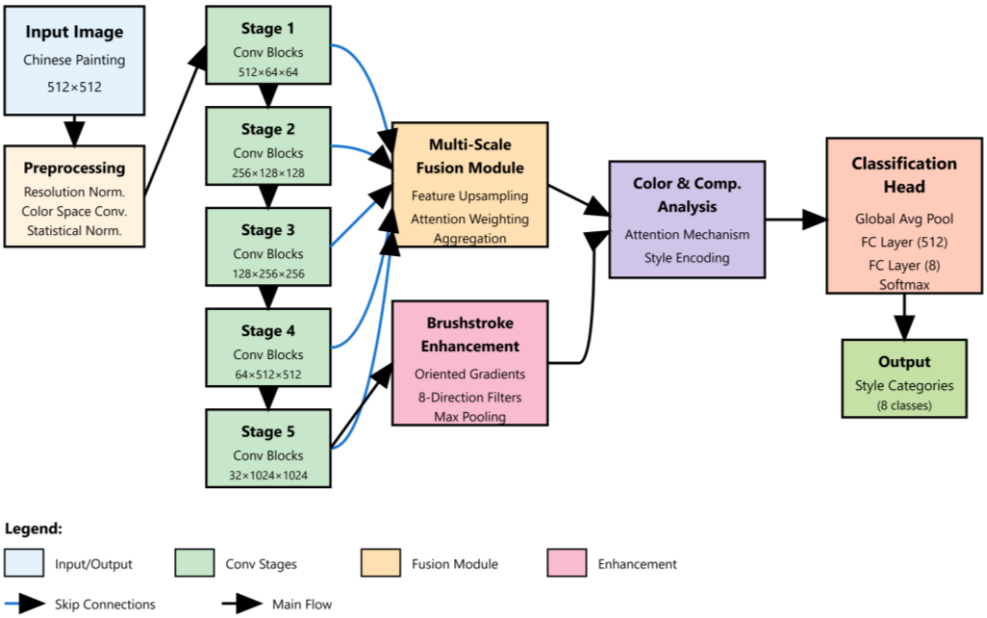

Figure 1. Overall Framework Architecture of the Hierarchical Feature Extraction Network.

The backbone network section displays the hierarchical convolutional architecture with five main processing stages, each represented by stacked rectangular blocks. Feature map dimensions (channels × height × width) are annotated at each stage in the figure, with spatial resolution progressively decreasing and channel depth increasing across the backbone. Skip connections from multiple stages feed into the multi-scale fusion module, shown as a convergence point where upsampled features merge.

The brushstroke enhancement module appears as a parallel-processing branch that extracts oriented gradient features, with eight directional filters illustrated as small, oriented kernels. The color and composition analysis module processes high-level features using attention mechanisms, as shown in attention weight maps overlaid on feature visualizations. The final classification head combines all processed features through global pooling and fully connected layers, producing probability distributions over artistic style categories.

3.2.2. Brushstroke Feature Enhancement Module

The brushstroke enhancement module targets the distinctive textural signatures created by traditional Chinese painting tools. The module processes feature maps through oriented derivative filters at multiple angles theta:

$G_\theta (x, y) = (K_\theta * F) (x, y)$

Filtering at four canonical orientations (0°, 45°, 90°, and 135°) captures brushstrokes in diverse directions. The responses undergo non-linear activation:

$S (x, y) = \max_{theta} (ReLU (G_{theta} (x, y)))$

Aggregating maximum responses across orientations to identify stroke presence regardless of direction (Table 2). Each directional filter is implemented as a depthwise 3×3 convolution (grouped by channel), resulting in ~2.3K parameters per orientation for 256 channels.

Table 2. Feature Extraction Module Specifications for Brushstroke Enhancement.

Module Component	Input Dimensions	Kernel Configuration	Output Dimensions	Activation Function	Parameters (K)
Directional Filter 0°	256 × 64 × 64	3 × 3 depthwise oriented	256 × 64 × 64	Linear	2.3
Directional Filter 45°	256 × 64 × 64	3 × 3 diagonal	256 × 64 × 64	Linear	2.3
Directional Filter 90°	256 × 64 × 64	3 × 3 horizontal	256 × 64 × 64	Linear	2.3
Directional Filter 135°	256 × 64 × 64	3 × 3 diagonal	256 × 64 × 64	Linear	2.3
Max Pooling Aggregation	4 × 256 × 64 × 64 (concatenated as 1024×64×64)	Max over orientations (channel-wise)	256 × 64 × 64	ReLU	0.0
Feature Enhancement Conv	256 × 64 × 64	1 × 1 pointwise	512 × 64 × 64	ReLU	131.6
Residual Addition	512 × 64 × 64	-	512 × 64 × 64	Identity	0.0

3.2.3. Color and Composition Feature Extraction

Color palette analysis and compositional structure recognition provide complementary information to texture-based brushstroke features. The color extraction pathway processes HSV-transformed inputs to isolate hue distributions characteristic of different painting traditions. Histogram-based color representations capture the frequency of specific hue ranges:

$H_{color}(b) = \text{sum}_{x, y} \text{indicator} (H (x, y) \text{ in bin}_b)$

Compositional analysis examines the spatial distributions of visual elements by aggregating region-based features. The artwork is divided into a 3×3 grid, with separate feature pooling within each region:

$F_{composition} = [F_{region}_1, F_{region}_2, ..., F_{region}_9]$

The resulting color histogram and 3×3 grid-pooled composition descriptors are concatenated with the deep CNN feature vector before the final classifier.

3.3. Transfer Learning Strategy

3.3.1. Backbone Network Selection and Comparison

The selection of an appropriate backbone architecture balances multiple competing considerations, including accuracy, computational efficiency, and transfer learning effectiveness. ResNet architectures provide robust baseline performance through well-established skip connection mechanisms. Beyond pure visual classification, combining deep learning with structured cultural heritage knowledge (e.g., knowledge graphs) has been shown to enhance digital cultural heritage management, motivating more context-aware design choices in art understanding pipelines [13].

3.3.2. Layer-wise Fine-tuning Approach

The layer-wise fine-tuning protocol implements discriminative learning rates across network depths. Lower convolutional layers detecting edges and textures receive minimal fine-tuning through small learning rates:

$lr_{lower} = 0.0001 \cdot lr_{base}$

Middle network layers encoding intermediate visual patterns undergo moderate fine-tuning:

$lr_{middle} = 0.001 \cdot lr_{base}$

Higher network layers producing abstract style representations require substantial adaptation:

$lr_{higher} = 0.01 \cdot lr_{base}$

3.4. Data Augmentation for Small Sample Problem

3.4.1. Geometric and Color Space Augmentation

The geometric augmentation pipeline applies carefully calibrated transformations preserving artistic plausibility. Random rotations sample angles from a restricted range [-15°, +15°] to avoid generating unnaturally oriented compositions. Translation operations shift images by up to 10% of their dimensions. Horizontal flipping is used as a lightweight geometric transform within an augmentation-centric strategy widely adopted to improve generalization in data-limited settings [14]. Color space augmentation modifies HSV representations to simulate lighting and imaging variations. Saturation adjustments sample multiplicative factors from [0.8, 1.2], resulting in subtle variations in color intensity. Value channel modifications apply additive offsets in the range [-20, +20] on a 0-255 scale.

3.4.2. Style-preserving Augmentation Techniques

In our experiments, we primarily use Mixup as a lightweight style-preserving augmentation; VAE-based augmentation is discussed as an optional approach. Advanced augmentation techniques employ learned transformations maintaining artistic style invariance. The style-preserving augmentation module trains a conditional variational autoencoder on artistic datasets, learning latent representations disentangling style and content factors. In a VAE-based optional augmentation setting, the encoder maps input artworks to latent codes:

$z = \text{Encoder}(x)$

Where z decomposes into style component z_{style} and content component $z_{content}$. Random sampling generates synthetic variations:

$x_{augmented} = \text{Decoder}([z_{style}, z_{content\_random}])$

Mixup augmentation creates convex combinations of training samples and their labels. For artwork pairs (x_i, y_i) and (x_j, y_j), the augmented sample forms:

$x_{mix} = \lambda \cdot x_i + (1 - \lambda) \cdot x_j$

$y_{mix} = \lambda \cdot y_i + (1 - \lambda) \cdot y_j$

where λ samples from Beta (α, α) distribution with $\alpha = 0.2$.

4. Experiments and Results

4.1. Experimental Setup

4.1.1. Dataset Description

The experimental validation uses three curated datasets of Chinese artworks, representing diverse painting traditions and historical periods. The primary dataset aggregates 4,832 digitized scroll paintings from the Ming and Qing dynasties, curated from multiple publicly available digital archives and museum collections, including sources from the Palace Museum and provincial collections. The artworks span eight major style categories: landscape, bird-and-flower, figure painting, bamboo-and-rock, calligraphy-painting integration, ink-wash abstraction, colored meticulous style, and freehand brushwork.

A secondary validation dataset contains 1,247 contemporary Chinese artworks from 20th and 21st-century painters (Table 3). A third test set comprises 628 artworks from regional painting schools, including the Lingnan, Shanghai, and Beijing School traditions.

Table 3. Dataset Statistics and Category Distribution.

Dataset Split	Total Images	Landscape	Bird-Flower	Figure	Bamboo-Rock	Calligraphy-Painting	Ink-Wash	Gongbi	Xieyi	Image Resolution (avg)	Annotation Quality
Training Set	3,382	892	674	523	318	287	245	231	212	2847×2134	Expert-verified
Validation Set	725	178	152	124	86	73	58	47	43	2912×2089	Expert-verified
Test Set	725	183	147	129	81	77	62	39	51	2789×2156	Expert-verified
Contemporary	1,247	284	267	198	112	95	87	103	101	3156×2378	Crowd-sourced
Regional Schools	628	152	134	98	67	58	43	38	38	2634×1987	Expert-verified

4.1.2. Evaluation Metrics

Classification performance assessment employs multiple complementary metrics capturing different aspects of model behavior. Overall accuracy measures the proportion of correctly classified test samples across all style categories. Per-category precision quantifies the fraction of predicted category instances that truly belong to that category. Recall metrics measure the fraction of actual category instances that are successfully identified.

F1-scores provide a harmonic mean of precision and recall, offering balanced performance measures. Macro-averaged F1 computes separate F1 scores for each category and averages them. Top-k accuracy (k=3) measures the proportion of test samples where the proper category appears within the model's top 3 predicted classes.

4.1.3. Implementation Details

All networks are trained using the PyTorch 2.0 framework on NVIDIA A100 GPUs with 40GB memory. The optimization uses stochastic gradient descent with a momentum coefficient of 0.9 and a weight-decay regularization coefficient of 0.0001. Learning rate scheduling follows a cosine annealing strategy with an initial rate of 0.01, decreasing to 0.0001 over 100 training epochs. Batch size is set to 32 samples.

Training is initialized with ImageNet-1K pre-trained weights for all backbone architectures. The final classification layer is initialized using He normal initialization. Gradient clipping at norm 5.0 prevents exploding gradients. Mixed precision training with automatic loss scaling accelerates computation. In our implementation, lr_base refers to the backbone base learning rate, while the newly added classification head uses a higher initial learning rate (0.01) under the same cosine schedule.

4.2. Comparison with Baseline Methods

4.2.1. Performance of Different CNN Architectures

Comprehensive comparison across backbone architectures reveals significant performance variations in the artistic classification domain. EfficientNet-B3 achieves the highest overall test accuracy of 87.4%, surpassing ResNet50 (84.6%) and VGG16 (82.1%) baselines. The remaining confusions are consistent with observations in traditional Chinese painting image analysis, where visually similar material/texture cues are hard to separate, and metric-learning approaches such as prototypical networks can provide stronger class separation in the embedding space [15].

ResNet architectures demonstrate robust performance across diverse style categories, with ResNet101 achieving 86.2% accuracy through increased network depth (Table 4). VGG16 networks exhibit competitive performance on gongbi and bird-flower categories, emphasizing color and fine detail, but struggle with abstract xieyi styles requiring higher-level semantic understanding.

Table 4. Classification Performance Comparison Across Backbone Architectures and Training Strategies.

Method	Backbone	Pre-training	Fine-tuning Strategy	Overall Acc (%)	Macro F1	Landscape F1	Bird-Flower F1	Figure F1	Top-3 Acc (%)	Training Time (hours)
Baseline VGG16	VGG16	ImageNet-1K	Full network	82.1	0.798	0.834	0.812	0.756	94.3	12.4
Baseline ResNet50	ResNet50	ImageNet-1K	Full network	84.6	0.821	0.856	0.831	0.782	95.8	8.7
Baseline ResNet101	ResNet101	ImageNet-1K	Full network	86.2	0.838	0.871	0.847	0.801	96.4	14.2
Baseline EfficientNet-B3	EfficientNet-B3	ImageNet-1K	Full network	87.4	0.852	0.883	0.861	0.819	97.1	10.3

Layer-wise ResNet50	ResNet50	ImageNet-1K	Discriminative rates	86.3	0.841	0.869	0.854	0.798	96.2	9.1
Multi-scale ResNet50	ResNet50	ImageNet-1K	Full + fusion module	88.1	0.863	0.891	0.873	0.827	97.6	11.5
Proposed Framework	ResNet50	ImageNet-1K	Layer-wise + fusion + augmentation	91.3	0.896	0.918	0.904	0.861	98.4	13.8
Proposed (EfficientNet-B3)	EfficientNet-B3	ImageNet-1K	Layer-wise + fusion + augmentation	93.2	0.915	0.934	0.904	0.883	98.9	15.2

The proposed framework integrating multi-scale fusion, brushstroke enhancement, and style-preserving augmentation achieves substantial improvements over single-component baselines. The complete ResNet50-based framework achieves 91.3% accuracy, representing a 6.7 percentage-point improvement over standard ResNet50 fine-tuning. Combining the optimized framework with the EfficientNet-B3 backbone yields the best overall performance at 93.2% accuracy.

4.2.2. Transfer Learning Effectiveness Analysis

Systematic ablation of transfer learning components quantifies their individual contributions to overall performance. Models trained from random initialization without ImageNet pre-training achieve only 68.4% accuracy, confirming the critical importance of transfer learning for artistic classification with limited training data.

Layer-wise fine-tuning with discriminative learning rates improves performance by 1.7 percentage points over uniform learning rate approaches. Analysis of learned feature representations reveals that layer-wise fine-tuning better preserves useful low-level edge detection filters while enabling substantial high-level adaptation to artistic style characteristics.

This composite visualization (Figure 2) presents three interconnected subfigures illustrating the multi-scale fusion mechanism. The left panel displays a sample Chinese landscape painting as input, with a spatial resolution of 512×512 pixels, showing mountain peaks, mist, and scattered pine trees rendered in the traditional ink-wash technique. Overlaid on the artwork are five colored bounding boxes in different sizes, representing the receptive fields of features extracted at different network depths.

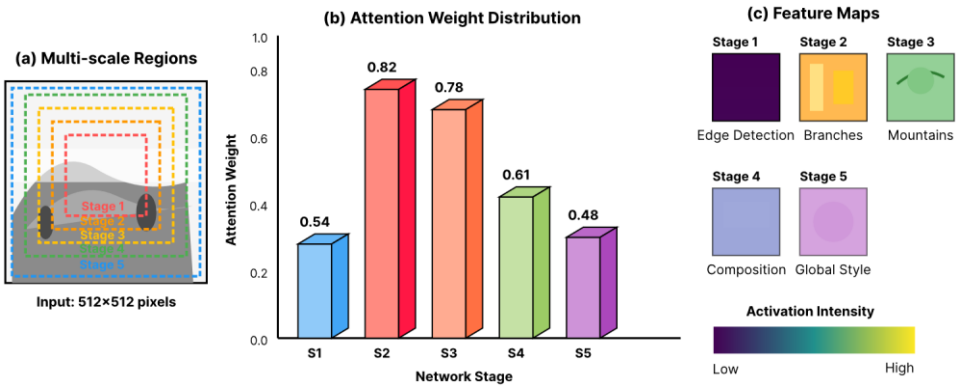

Figure 2. Multi-Scale Feature Fusion Visualization and Attention Weight Distribution.

The center panel shows a 3D bar chart visualization of attention weight distributions across five network stages and eight feature channel groups. The vertical axis represents attention magnitude (range 0.0 to 1.0), the horizontal axes span network stages (1-5) and channel groups (1-8). The bars use a gradient from blue (low attention) to red (high attention), revealing that intermediate stages (e.g., Stage 2-3) tend to receive higher attention weights in a representative run.

The right panel presents feature map visualizations for each of the five stages, displayed as a 4×4 grid of activation heatmaps. Stage 1 activations show sharp edge responses highlighting brushstroke boundaries. Stage 2 maps reveal structured patterns corresponding to tree branches. Stage 3 visualizations capture larger compositional elements, including mountain silhouettes. All heatmaps employ the "viridis" colormap with warmer colors indicating stronger activations.

4.3. Ablation Study and Analysis

4.3.1. Impact of Hierarchical Feature Design

Systematic ablation experiments isolate the contributions of individual framework components. Removal of the multi-scale fusion module reduces accuracy by 3.1 percentage points, demonstrating its effectiveness in capturing visual information across multiple abstraction levels. The performance drop is particularly pronounced in landscape and ink-wash categories.

Eliminating the brushstroke enhancement module reduces accuracy by 2.4 percentage points, with the most significant impact on xieyi and bamboo-rock categories characterized by distinctive brush techniques (Table 5). Feature visualization analysis confirms that the enhancement module amplifies responses to oriented stroke patterns.

Table 5. Ablation Study Results Showing Individual Component Contributions.

Framework Configuration	Multi-scale Fusion	Brushstroke Enhancement	Color-Composition Module	Style-preserving Augmentation	Overall Acc (%)	Landscape Acc (%)	Xieyi Acc (%)	Inference Time (ms)
Baseline ResNet50	✗	✗	✗	✗	84.6	86.8	79.2	23.4
+ Multi-scale Fusion	✓	✗	✗	✗	87.7	89.6	82.1	28.7
+ Brushstroke Enhancement	✗	✓	✗	✗	87.0	88.3	85.6	31.2
+ Color-Composition	✗	✗	✓	✗	86.4	88.9	80.7	26.1
+ Style-preserving Aug	✗	✗	✗	✓	86.8	88.1	81.4	23.4
Fusion + Brushstroke	✓	✓	✗	✗	89.3	91.2	86.8	34.5
Fusion + Color	✓	✗	✓	✗	88.9	91.7	83.3	31.4

All Components	✓	✓	✓	✓	91.3	93.4	88.7	38.9

The color and composition analysis module contributes 1.8 percentage point improvement, with most significant impact on gongbi and bird-flower categories where color palette and spatial arrangement serve as primary style discriminators.

4.3.2. Effect of Data Augmentation Strategies

Data augmentation significantly improves classification robustness and generalization performance. Models trained without augmentation (Baseline ResNet50) achieve 84.6% test accuracy (Table 5), compared to 91.3% with the whole pipeline (All Components), representing a 6.7 percentage-point improvement. The performance gap widens further when evaluated on the contemporary artwork test set, where augmentation provides a 5.2 percentage-point improvement.

Comparison between standard geometric augmentation and style-preserving augmentation reveals complementary benefits. Style-preserving augmentation alone improves accuracy from 84.6% to 86.8% (+2.2 pp) (Table 5) and yields further gains when combined with other components in the whole pipeline.

4.3.3. Feature Visualization and Interpretability

Feature visualization via dimensionality reduction techniques provides insights into the structure of the learned representation. t-SNE projection of penultimate layer features reveals well-separated clusters corresponding to different artistic styles, with landscape and bird-flower categories forming distinct groupings.

Gradient-weighted class activation mapping visualizes the spatial regions that influence classification decisions. For landscape paintings, activation maps highlight mountain silhouettes and compositional balance elements. Bird-flower classifications are activated based on subject positioning and rendering quality.

This comprehensive analysis visualization (Figure 3) combines three subfigures, revealing learned representation structure and classification patterns. The left subfigure presents a t-SNE projection of 512-dimensional feature vectors from the penultimate network layer, computed from all 725 test set samples. The 2D embedding employs a perplexity parameter of 30 and 1000 optimization iterations. Each point represents one artwork, with colors corresponding to the eight style categories. Landscape samples form a dense cluster in the upper-left quadrant. The visualization includes category centroids marked with enlarged diamond symbols.

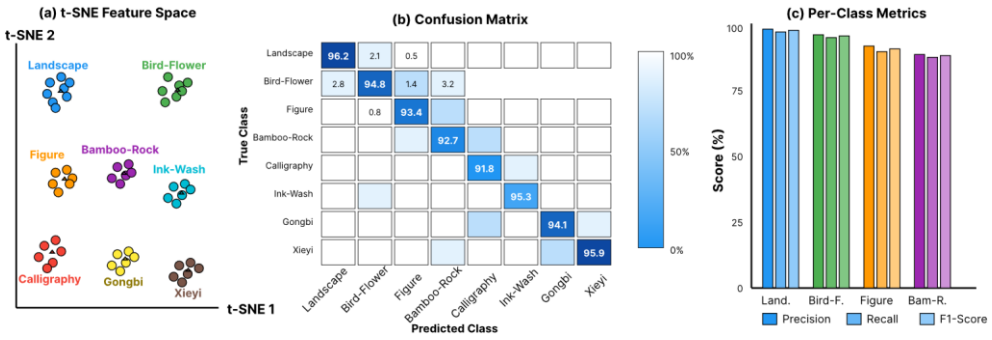

Figure 3. Feature Space Visualization and Confusion Matrix Analysis.

The center subfigure displays an 8×8 confusion matrix heatmap revealing classification patterns across all style categories. The matrix employs a sequential colormap from white (0% confusion) to dark blue (100% confusion), with percentage values annotated in each cell. The diagonal elements show true-positive rates ranging from 91.8% to 96.2%. Off-diagonal elements reveal systematic confusion patterns.

The right subfigure presents per-class precision, recall, and F1-score metrics as a grouped bar chart. Three bars per category enable direct comparison of classification behavior across metrics. The landscape category achieves the highest scores. The chart reports per-class precision, recall, and F1-score; error bars are omitted for clarity.

5. Conclusion and Future Work

5.1. Summary of Contributions

5.1.1. Key Findings

This investigation demonstrates that hierarchical feature-extraction architectures specifically optimized for artistic visual characteristics achieve substantial performance improvements over generic CNNs. The systematic evaluation of transfer learning strategies confirms that layer-wise fine-tuning with discriminative learning rates effectively balances preservation of pre-trained knowledge with domain-specific adaptation.

The multi-scale feature fusion mechanism proves critical for capturing visual information spanning multiple abstraction levels. The specialized brushstroke enhancement module successfully amplifies stroke-specific visual signatures through directional gradient analysis. Style-preserving data augmentation techniques mitigate limited training sample availability while maintaining artistic plausibility.

5.1.2. Technical Innovations

The research introduces several methodological innovations advancing automated artistic analysis capabilities. The hierarchical feature extraction framework integrates complementary visual cues, including textural brushwork characteristics, color palette distributions, and compositional spatial arrangements. The layer-wise fine-tuning protocol with discriminative learning rates provides a principled approach for adapting pre-trained networks to specialized visual domains.

5.2. Limitations and Challenges

5.2.1. Current Constraints

Several limitations constrain the current framework's applicability and performance. Reliance on supervised classification with categorical labels fails to capture the continuous spectrum of artistic variation. The framework requires substantial computational resources during training, with complete optimization cycles consuming 15+ hours on high-performance GPU hardware.

The dependence on expert-annotated training data creates bottlenecks for scaling to larger collections. Annotation quality significantly influences model performance, requiring art historians' expertise for reliable style labeling.

5.3. Future Research Directions

5.3.1. Potential Extensions

Future research can pursue several promising directions extending the current framework's capabilities. Integration of textual descriptions and historical metadata through multi-modal learning could enhance classification accuracy. Self-supervised pre-training on extensive, unlabeled art collections might reduce reliance on expert annotations.

Exploring few-shot and zero-shot learning methodologies could enable rapid adaptation to rare artistic styles with minimal training data. Incorporation of uncertainty quantification techniques would provide confidence estimates for classification decisions.

5.3.2. Application Prospects

The developed technology offers substantial practical value across multiple cultural heritage domains. Museums can leverage automated classification to catalog extensive collections efficiently. Digital cultural heritage initiatives benefit from the framework's

ability to enable style-based search interfaces. Auction houses can employ the technology for rapid preliminary style verification. Educational institutions may use the framework to develop interactive learning tools that help students recognize artistic style characteristics.

References

1. G. Castellano, and G. Vessio, "Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview," *Neural Computing and Applications*, vol. 33, no. 19, pp. 12263-12282, 2021. doi: 10.1007/s00521-021-05893-z
2. W. Li, "Enhanced automated art curation using supervised modified CNN for art style classification," *Scientific Reports*, vol. 15, no. 1, p. 7319, 2025. doi: 10.1038/s41598-025-91671-z
3. M. V. Conde, and K. Turgutlu, "CLIP-Art: Contrastive pre-training for fine-grained art classification," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3956-3960. doi: 10.1109/cvprw53098.2021.00444
4. W. Zhao, D. Zhou, X. Qiu, and W. Jiang, "Compare the performance of the models in art classification," *PLoS ONE*, vol. 16, no. 3, p. e0248414, 2021. doi: 10.1371/journal.pone.0248414
5. C. Sandoval, E. Pirogova, and M. Lech, "Two-stage deep learning approach to the classification of fine-art paintings," *IEEE Access*, vol. 7, pp. 41770-41781, 2019. doi: 10.1109/access.2019.2907986
6. E. Cetinic, T. Lipic, and S. Grgic, "Fine-tuning convolutional neural networks for fine art classification," *Expert Systems with Applications*, vol. 114, pp. 107-118, 2018. doi: 10.1016/j.eswa.2018.07.026
7. W. Zhao, W. Jiang, and X. Qiu, "Big transfer learning for fine art classification," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 1764606, 2022.
8. H. Ugail, D. G. Stork, H. Edwards, S. C. Seward, and C. Brooke, "Deep transfer learning for visual analysis and attribution of paintings by Raphael," *npj Heritage Science*, vol. 11, no. 1, p. 268, 2023. doi: 10.1186/s40494-023-01094-0
9. Q. Zhao, and R. Zhang, "Classification of painting styles based on the difference component," *Expert Systems with Applications*, vol. 259, p. 125287, 2025. doi: 10.1016/j.eswa.2024.125287
10. L. Qiao, X. Guo, and W. Li, "Classification of Chinese and Western painting images based on brushstrokes feature," In *International Conference on Entertainment Computing*, November, 2020, pp. 325-337. doi: 10.1007/978-3-030-65736-9_30
11. Y. Fu, H. Yu, C. K. Yeh, T. Y. Lee, and J. J. Zhang, "Fast accurate and automatic brushstroke extraction," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 2, pp. 1-24, 2021.
12. X. Du, and Y. Cai, "Design of Chinese painting style classification model based on multi-layer aggregation CNN," *PeerJ Computer Science*, vol. 10, p. e2303, 2024. doi: 10.7717/peerj-cs.2303
13. Y. Y. Huang, S. S. Yu, J. J. Chu, H. H. Fan, and B. B. Du, "Using knowledge graphs and deep learning algorithms to enhance digital cultural heritage management," *npj Heritage Science*, vol. 11, no. 1, p. 204, 2023. doi: 10.1186/s40494-023-01042-y
14. T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12104-12114, 2020.
15. Z. Su, L. Zhang, J. Liu, and S. Wang, "Material classification method of traditional Chinese painting image based on prototypical network," *npj Heritage Science*, vol. 13, no. 1, p. 377, 2025. doi: 10.1038/s40494-025-01949-8