

Lightweight Artistic Style Classification via Texture Correlation and Adaptive Local-Global Fusion

Jinjiang Zhao^{1, 2, a, *}, Peng Ren², Rui Guo¹, Lei Zhang^{1, b, *}

¹ Institute of Automation, Gansu Academy of Sciences, Lanzhou, 730000, Gansu, China

² College of Automation and Electrical Engineering, Lanzhou University of Technology, Lanzhou 730050, Gansu, China

^a red eggs85@163.com, ^b hanssomezjj@126.com

Abstract: Artistic style classification remains challenging because visually similar styles often exhibit subtle differences in brushstrokes, textures, colors, and global composition. To address this problem, this paper proposes a lightweight artistic style classification framework based on a MobileNet-Transformer Hybrid (MTH). First, a Lightweight Texture Correlation Module (LTCM) is introduced to capture multi-scale texture relationships from intermediate convolutional features. Second, an Adaptive Local-Global-Texture Fusion module (ALGTF) dynamically integrates local details, global composition, and texture representations according to their relative contributions. Third, a Confusable-Style Hard Contrastive Learning strategy (CSHCL) is designed to emphasize difficult inter-style samples and improve feature separability between visually similar artistic categories. The Dynamic Channel-Spatial Attention Module (DCSAM) is retained to enhance style-sensitive regions and channels. Experiments on ArtBench-10 and Painting91_Styles demonstrate that the proposed framework provides improved classification performance while maintaining a lightweight architecture, showing its effectiveness for fine-grained artistic image analysis.

Keywords: Artistic style classification, lightweight deep learning, texture correlation, adaptive feature fusion, contrastive learning.

1. Introduction

The rapid digitization of cultural collections has created large-scale repositories of paintings and other visual artworks, making automatic artistic style classification increasingly important for digital archiving, artwork retrieval, museum collection management, and intelligent recommendation. Unlike general image classification, artistic style recognition depends less on object semantics and more on subtle visual characteristics, including brushstroke patterns, texture statistics, color organization, and global composition. These characteristics are often distributed across different spatial scales, making fine-grained style discrimination particularly challenging when visually similar artistic movements share comparable subjects or color palettes.

Recent deep learning approaches have substantially improved artistic image analysis. Convolutional Neural Networks (CNNs) are effective at capturing local textures, edges, and brushstroke details, but their limited receptive fields make long-range compositional relationships difficult to model. In contrast, Vision Transformers (ViTs) provide strong global dependency modeling through self-attention, but their computational cost and weaker local inductive bias can limit their effectiveness in lightweight fine-grained classification. To combine these complementary properties, Ma et al. [1] proposed MTH, which integrates depthwise separable convolution with a Transformer branch and further introduces DCSAM and cross-style contrastive learning. The model achieves a favorable balance between classification performance and computational efficiency, making it an appropriate lightweight foundation for artistic image classification.

However, artistic style is not fully represented by conventional convolutional features alone. Texture correlation is particularly important because brushstroke organization and repeated local patterns frequently

distinguish visually similar styles. Wang and Feng [2] proposed MSA-Net, in which multi-scale texture information is modeled using group-wise Gram aggregation and learnable Gram representations, while structural and color information are extracted through additional branches. Their results demonstrate that explicitly modeling correlations among feature channels can improve style representation beyond standard deep features. Nevertheless, directly introducing the complete multi-stream architecture into a lightweight network would substantially increase structural complexity.

Another important issue is how local and global representations are combined. Simple concatenation or addition assigns relatively fixed importance to different feature sources and cannot adapt to the visual characteristics of individual artworks. ArtFusionNet [3] addresses this limitation through adaptive fusion between multi-scale CNN features and Transformer representations, allowing the contribution of local and global information to be dynamically adjusted. Although this strategy improves feature integration, the model relies on a relatively heavy ResNet-based backbone together with dilated convolution, pyramid pooling, and Transformer encoding, which increases the overall computational burden.

In addition, style confusion remains a major difficulty in fine-grained artistic image classification. Conventional contrastive learning can improve inter-class separability, but randomly selected negative samples are often already visually dissimilar and therefore provide limited discriminative supervision. Recent work has shown that contrastive learning is particularly useful for separating difficult artistic styles and suppressing irrelevant visual responses [4]. This suggests that explicitly emphasizing hard negative samples from visually confusable categories may provide stronger supervision without increasing inference complexity.

Motivated by these observations, this study develops an improved lightweight artistic style classification framework

based on MTH. The proposed model retains the efficient MobileNet-based backbone and global Transformer branch while introducing three targeted modifications. First, LTCM extracts multi-scale texture correlations from intermediate convolutional features using channel grouping and lightweight learnable correlation modeling. Second, ALGTF dynamically integrates local, global, and texture features using learned fusion weights, allowing the network to adapt feature importance to different artistic styles. Third, CSHCL selects visually similar inter-class samples as hard negatives during training, thereby strengthening the separation of confusable style categories while introducing no additional inference branch. The fused representation is subsequently refined by DCSAM before classification.

The proposed framework is evaluated on ArtBench-10 and Painting91_Styles. The former provides a relatively large and balanced benchmark for evaluating overall style classification performance, whereas the latter contains 13 fine-art style categories and is suitable for assessing discrimination among more challenging and visually related artistic styles [1, 2]. Together, the two datasets provide complementary evaluation conditions for examining classification accuracy, generalization, and fine-grained style representation.

The main contributions of this work are summarized as follows:

(1) A lightweight multi-scale texture modeling strategy is introduced to explicitly capture channel-wise texture correlations from intermediate backbone features. Compared with directly adopting a heavy multi-stream architecture, LTCM preserves the lightweight characteristic of MTH while strengthening art-specific texture representation.

(2) An adaptive feature fusion mechanism is designed to dynamically integrate local details, global composition, and texture correlations. The learned weighting strategy enables the network to adjust the contribution of different feature sources according to the visual characteristics of each artwork.

(3) A hard contrastive learning strategy is proposed for visually confusable artistic styles. By prioritizing difficult negative samples in the feature space, CSHCL enhances inter-style separability during training without increasing the computational cost of inference.

The remainder of this paper is organized as follows. Section 2 reviews recent studies on artistic image classification, texture representation, and hybrid feature modeling. Section 3 presents the proposed framework and its optimization objective. Section 4 describes the experimental setup and analyzes the results on the two artistic image datasets. Section 5 concludes the paper and discusses future work.

2. Related Works

2.1. Deep Learning for Artistic Style Classification

Artistic style classification differs from conventional object recognition because stylistic information is distributed across brushstrokes, textures, color arrangements, and global composition. Recent studies have therefore moved from single CNN architectures toward hybrid representations capable of modeling both local and global visual characteristics. Ma et al. [1] proposed MTH, which combines MobileNetV3 with a lightweight Transformer branch. The CNN component captures local style details, while self-attention models long-range dependencies. DCSAM is further

introduced to enhance style-sensitive channels and spatial regions, and CSFT improves feature discrimination through contrastive learning. This lightweight design provides the backbone adopted in the present study.

Hybrid CNN-Transformer architectures have also been investigated for more complex style representation. ArtFusionNet [3] employs a modified ResNet-50 with dilated convolution and pyramid pooling to extract multi-scale local features, while a Transformer models global contextual relationships. An AFM dynamically balances the contributions of CNN and Transformer representations instead of using fixed feature concatenation. The results indicate that adaptive local-global integration is beneficial for fine-grained artistic style recognition, although the resulting architecture is relatively computationally intensive.

Other recent methods have explored alternative feature extraction strategies. Xiao and Li [5] combine Xception with PCA to reduce high-dimensional redundancy while retaining discriminative painting features. Xu and Lu [6] separately extract local and global characteristics from artistic images and employ attention-based fusion before classification. These studies further demonstrate that accurate artistic style recognition requires complementary information from different feature levels rather than relying on a single visual representation.

2.2. Texture and Color Representation

Texture is one of the most distinctive characteristics of paintings because different artistic movements often exhibit characteristic brushstroke distributions and surface patterns. Wang and Feng [8] proposed MSA-Net to explicitly model texture, structural, and color information. Its texture stream applies group-wise Gram aggregation to multi-level ResNet features and introduces L-Gram to reduce redundant correlation information. A Transformer is subsequently used to refine the aggregated texture representation. The structural stream employs global average pooling, while the color stream describes RGB distributions. Their experiments show that combining these complementary style modalities provides stronger representations than texture-only modeling. Recent research has also emphasized the importance of chromatic information. Zhang [7] proposed a color-perception-aware framework that integrates color enhancement and transformations in HSV and Lab spaces with a ResNet-based classifier. The results on WikiArt indicate that explicit color modeling can complement structural and textural representations. However, directly introducing multiple independent streams or handcrafted color transformations can increase model complexity. Therefore, the present work focuses on a lightweight texture correlation branch extracted directly from intermediate backbone features, preserving the efficiency of the original MTH architecture.

2.3. Feature Fusion and Discriminative Style Learning

The manner in which different feature sources are combined is another important factor in artistic image classification. Simple addition or concatenation treats different representations relatively uniformly and may not reflect the visual characteristics of individual artworks. ArtFusionNet [3] addresses this issue using learnable adaptive weighting between local CNN features and global Transformer features. Inspired by this strategy, the proposed

method further extends adaptive fusion to three complementary representations: local details, global composition, and texture correlation. This enables the contribution of each representation to be dynamically adjusted during classification.

In addition to feature fusion, recent work has shown that contrastive learning can improve the separation of visually similar artistic styles. The original CSFT strategy [1] introduces contrastive constraints to improve cross-style robustness, while more recent artistic style classification methods combine contrastive objectives with attention mechanisms to suppress irrelevant regions and strengthen discriminative features [4]. Nevertheless, randomly selected negative samples may provide limited supervision when they are already visually distinct from the anchor image. The proposed CSHCL therefore emphasizes hard negative samples from confusable style categories, allowing the model to learn more discriminative decision boundaries without increasing inference complexity.

Overall, existing studies have demonstrated the effectiveness of lightweight hybrid networks, texture correlation modeling, adaptive local-global fusion, and contrastive learning. However, these techniques are generally investigated separately. The proposed method integrates lightweight multi-scale texture modeling, adaptive local-global-texture fusion, and confusable-style-oriented contrastive learning within a unified MTH-based framework.

3. Proposed Method

3.1. Overall Framework

The overall architecture of the proposed framework is

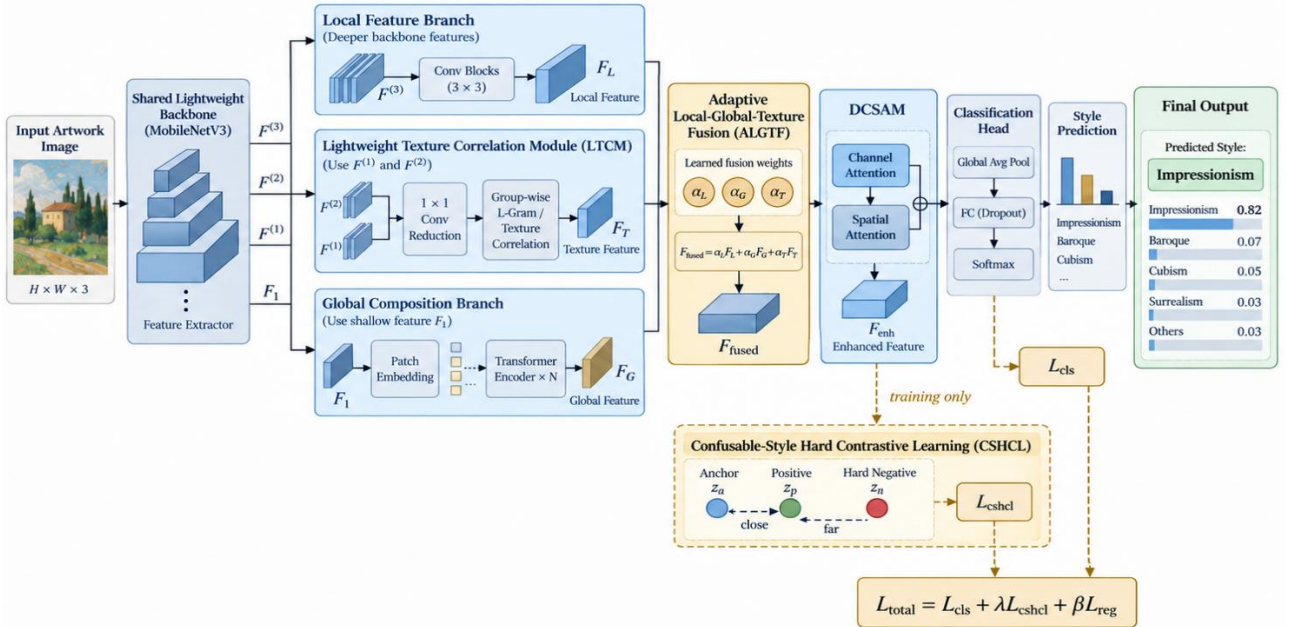


Figure 1. Overall architecture of the proposed art style classification network

Convolution is adopted to reduce computational cost while preserving local feature extraction capability:

$$F_{local} = \text{Conv}_{1 \times 1}(\text{Conv}_{dw}^{k \times k}(X)) \quad (1)$$

Where X denotes the input feature map. Conv_{dw} denotes depthwise convolution, and $\text{Conv}_{1 \times 1}$ denotes pointwise

shown in Fig. 1. The model is developed on the basis of the lightweight MobileNet-Transformer Hybrid architecture [14] and consists of five major components: a shared lightweight backbone, a global composition branch, a Lightweight Texture Correlation Module (LTCM), an Adaptive Local-Global-Texture Fusion module (ALGTF), and a Dynamic Channel-Spatial Attention Module (DCSAM). During training, an additional Confusable-Style Hard Contrastive Learning strategy (CSHCL) is introduced to improve the separation of visually similar artistic styles. Given an input artwork image, MobileNetV3 first extracts hierarchical local features. Shallow features are fed into the Transformer branch to model global compositional dependencies, while intermediate features are delivered to LTCM to explicitly capture texture correlations. The local, global, and texture representations are subsequently integrated through ALGTF using learnable weights. DCSAM further enhances discriminative channels and spatial regions before the classification head produces the final style probability distribution. CSHCL is activated only during training and therefore does not introduce additional inference cost.

3.2. Training-Stage Residual Diffusion Prior

The shared backbone follows the lightweight design of MTH [14]. MobileNetV3 is employed to extract local style features, including brushstrokes, edges, and texture patterns. Depthwise separable.

convolution. F_{local} represents the local feature generated by the lightweight convolutional branch. Although convolution is effective for local pattern extraction, artistic composition often contains long-range dependencies. Therefore, a Transformer branch is applied to shallow backbone features. Query, key, value, and the corresponding attention output are calculated as:

$$Q_i = F_1 W_i^Q, \quad K_i = F_1 W_i^K, \quad V_i = F_1 W_i^V, \quad A_i = \text{Softmax} \left(\frac{Q_i K_i^T}{\sqrt{d_k}} \right) V_i \quad (2)$$

Where F_1 denotes the shallow backbone feature. W_Q , W_K , and W_V are learnable projection matrices. d_k denotes the dimensionality of the key vector, and A_i represents the output of the i -th attention head.

The outputs of multiple attention heads are concatenated to obtain the global representation:

$$F_G = \text{Concat}(A_1, A_2, \dots, A_N) W^O \quad (3)$$

Where N denotes the number of attention heads. W_O is the output projection matrix, and F_G denotes the global composition feature.

The local branch and Transformer branch therefore capture complementary information. The former focuses on fine-grained details, while the latter models long-range relationships among different regions of the artwork.

3.3. Lightweight Texture Correlation Module

Although MTH combines local and global information, conventional convolutional representations do not explicitly describe channel correlations associated with artistic textures. Inspired by the multi-scale Gram representation in MSA-Net [2], LTCM is introduced to model style-sensitive texture relationships while maintaining a lightweight structure.

Intermediate features from two MobileNet stages are first compressed using 1×1 convolution:

$$\hat{F}^{(l)} = \text{Conv}_{1 \times 1}(F^{(l)}) \quad (4)$$

Where F^l denotes the intermediate feature at the l -th backbone stage. $\hat{F}_{\text{hat}}$ denotes the channel-reduced feature used for texture modeling.

Each reduced feature is divided into several channel groups. For the g -th group, a correlation matrix is calculated as:

$$G_{l,g} = \frac{1}{N_{l,g}} \hat{F}_{l,g} \hat{F}_{l,g}^T \quad (5)$$

Where $\hat{F}_{l,g}$ denotes the feature representation of the g -th channel group. $N_{l,g}$ is the normalization factor determined by the spatial size of the feature map. $G_{l,g}$ represents the corresponding texture correlation matrix.

To reduce redundant information, only the lower triangular part of the correlation matrix is retained and transformed by a learnable projection:

$$L_{l,g} = L_2 \left(\text{tril}(G_{l,g}) W_g \right) \quad (6)$$

Where tril denotes the lower-triangular extraction operation. W_g is a learnable transformation matrix, and L_2 denotes L_2 normalization. $L_{l,g}$ represents the lightweight learnable texture correlation feature.

Texture features obtained from different scales and groups are aggregated and processed by a lightweight Transformer:

$$M_s = \sigma \left(\text{DWConv}_{3 \times 3}(\text{Concat}(\text{AvgPool}(F_{\text{fused}}), \text{MaxPool}(F_{\text{fused}}))) \right) \quad (11)$$

Where AvgPool and MaxPool compress the channel dimension using average and maximum operations. DWConv

$$F_T = \text{Transformer}(\text{Concat}(L_{1,1}, \dots, L_{1,G}, L_{2,1}, \dots, L_{2,G})) \quad (7)$$

Where F_T denotes the final texture representation. By explicitly modeling multi-scale channel correlations, LTCM provides complementary information that is difficult to capture using conventional convolutional features alone [17].

Compared with the complete three-stream structure of MSA-Net [2], the proposed LTCM uses only selected intermediate backbone features and avoids independent structural and color branches. Therefore, it strengthens texture representation without substantially increasing the complexity of the lightweight backbone.

3.4. Adaptive Local-Global-Texture Fusion and Attention Enhancement

The three extracted representations contain different types of style information. F_L describes local visual details, F_G represents global composition, and F_T captures texture correlations. Direct addition or concatenation cannot dynamically determine which representation is more important for a particular artwork. Inspired by the adaptive fusion strategy of ArtFusionNet, ALGTF learns the relative contribution of the three feature sources.

The fusion weights are calculated as:

$$[\alpha_L, \alpha_G, \alpha_T] = \text{Softmax}(\text{MLP}([F_L; F_G; F_T])) \quad (8)$$

Where α_L , α_G , and α_T denote the learned importance weights of the local, global, and texture features, respectively. The semicolon indicates feature concatenation before weight estimation.

The fused style representation is then obtained as

$$F_{\text{fused}} = \alpha_L F_L + \alpha_G F_G + \alpha_T F_T \quad (9)$$

Where F_{fused} denotes the adaptively fused feature. The normalized fusion weights allow the network to automatically adjust the contribution of different feature sources according to the characteristics of each artwork.

For example, texture information may receive a larger weight for styles dominated by characteristic brushstrokes, whereas global composition may contribute more strongly to geometrically structured styles.

After feature fusion, DCSAM is retained from the original lightweight model to enhance style-sensitive information. Channel attention is calculated as

$$M_c = \sigma(W_2 \delta(W_1 \text{GAP}(F_{\text{fused}}))) \quad (10)$$

Where GAP denotes global average pooling. W_1 and W_2 are learnable fully connected transformations. δ represents the ReLU activation function, σ denotes the Sigmoid function, and M_c is the channel attention map.

Spatial attention is calculated by combining channel-wise average and maximum pooling:

denotes depthwise convolution, and M_s represents the spatial attention map. The final enhanced feature is calculated as

$$F_{enh} = F_{fused} \odot M_c + F_{fused} \odot M_s \quad (12)$$

Where the $\odot$ symbol denotes element-wise multiplication. F_{enh} represents the enhanced style feature used for final classification.

ALGTF and DCSAM operate at different levels. ALGTF determines the importance of different feature sources, whereas DCSAM determines the importance of channels and spatial locations within the fused representation.

3.5. Style Classification and Confusable-Style Hard Contrastive Learning

The enhanced representation is first processed by global average pooling and a fully connected classification layer.

$$\left\{ \begin{aligned} n^* &= \arg \max_{y_n \neq y_a} \text{sim}(z_a, z_n), \\ L_{cshcl} &= -\log \frac{\exp(\text{sim}(z_a, z_p) / \tau)}{\exp(\text{sim}(z_a, z_p) / \tau) + \sum_{n \in H_a} \exp(\text{sim}(z_a, z_n) / \tau)} \end{aligned} \right\} \quad (14)$$

Where z_a denotes the anchor feature. z_p denotes a positive feature from the same artistic style. z_n denotes a negative feature from a different style. n_{star} represents the hardest negative sample. H_a denotes the selected hard-negative set. The function sim represents cosine similarity, and τ is the temperature coefficient.

This strategy forces the model to focus on difficult style boundaries, such as Impressionism versus Post-Impressionism or Baroque versus closely related classical styles. Since hard-negative mining is used only during training, it does not increase the computational cost of inference. The complete training objective is defined as:

$$L_{total} = L_{cls} + \lambda L_{cshcl} + \beta L_{reg} \quad (15)$$

Where L_{cls} denotes the standard cross-entropy classification loss. L_{cshcl} represents the proposed hard contrastive loss. L_{reg} denotes L2 regularization. λ controls the contribution of contrastive learning, while β controls the regularization strength.

The resulting framework therefore performs style recognition through three complementary mechanisms: LTCM strengthens art-specific texture representation, ALGTF dynamically integrates local, global, and texture information, and CSHCL improves the separation of visually confusable artistic styles. Together, these modifications preserve the lightweight characteristics of the original MTH architecture while enhancing its capability for fine-grained artistic style classification.

4. Experimental Results and Analysis

4.1. Experimental Setup

Experiments were conducted on ArtBench-10 and Painting91_Styles. ArtBench-10 contains 60,000 artwork images from 10 balanced artistic styles. Following the setting in [1], 50,000 images were used for training, 5,000 for validation, and 5,000 for testing. Painting91_Styles contains 2,338 paintings from 13 artistic styles [2], and the dataset was divided into training, validation, and testing sets with a ratio of 70%, 15%, and 15%, respectively. All images were resized to 256×256 . Standard data augmentation, including random cropping, horizontal flipping, slight rotation, and color

The final style probability is calculated as:

$$p(y|x) = \text{Softmax}(W_o \text{GAP}(F_{enh}) + b_o) \quad (13)$$

Where W_o and b_o denote the parameters of the classification layer. The output represents the probability distribution over all artistic style categories, and the category with the highest probability is selected as the final prediction.

To improve discrimination between visually similar styles, CSHCL is introduced during training. Unlike conventional contrastive learning that randomly selects negative samples, the proposed strategy selects the most similar sample belonging to a different style as the hard negative. The hard negative and contrastive objective are defined as:

jittering, was applied during training [18].

The proposed model was implemented using PyTorch. Adam was employed as the optimizer with an initial learning rate of 0.001, a batch size of 128, and a maximum of 200 training epochs. The Transformer branch used eight attention heads with a feature dimension of 256. In LTCM, two intermediate backbone features were used for grouped texture correlation modeling. The temperature coefficient of CSHCL was set to 0.1, while the weights of the contrastive loss and regularization term were set to 0.5 and 0.01, respectively.

Five representative models, including EfficientNetV2-S [16], Swin-Tiny, ConvNeXt-T [17], MobileViT-S [18], and the original MTH [19], were selected for comparison. All methods were evaluated under the same experimental protocol. Top-1 Accuracy, Macro Precision, and Macro F1-score were adopted as the main evaluation metrics.

4.2. Comparison with Existing Methods

To evaluate the effectiveness of the proposed method, five representative models were compared on the ArtBench-10 dataset, including EfficientNetV2-S, Swin-Tiny, ConvNeXt-T, MobileViT-S, and the original MTH. The comparison results are presented in Table 1.

The proposed method achieves the highest performance across all three evaluation metrics. Compared with the original MTH, the proposed model improves Top-1 Accuracy from 85.2% to 87.1%, Macro Precision from 84.7% to 86.6%, and Macro F1-score from 84.8% to 86.8%. The improvement indicates that explicitly modeling texture correlations through LTCM provides more discriminative style information, while ALGTF effectively coordinates local, global, and texture representations. In addition, CSHCL further improves the separation of visually similar artistic styles. Compared with conventional CNN, Transformer, and lightweight hybrid architectures, the proposed framework therefore provides stronger fine-grained artistic style classification capability.

Table 1. Comparison of classification performance on the ArtBench-10 dataset (%)

Method	Top-1 Accuracy	Macro Precision	Macro F1-score
EfficientNetV2-S [16]	78.9	78.2	78.5
Swin-Tiny [17]	80.6	80.1	80.3
ConvNeXt-T [18]	81.3	80.9	81.0
MobileViT-S [19]	79.7	79.3	79.5
Original MTH [1]	85.2	84.7	84.8
Proposed	87.1	86.6	86.8

Table 2. Ablation study of the proposed modules on ArtBench-10 (%)

Configuration	LTCM	ALGTF	CSHCL	Top-1 Accuracy	Macro Precision	Macro F1-score
Original MTH	—	—	—	85.2	84.7	84.8
MTH + LTCM	✓	—	—	85.9	85.4	85.6
MTH + LTCM + ALGTF	✓	✓	—	86.5	86.0	86.2
MTH + LTCM + CSHCL	✓	—	✓	86.3	85.9	86.1
Full model	✓	✓	✓	87.1	86.6	86.8

The results demonstrate that all three proposed modules contribute positively to artistic style classification. Introducing LTCM improves Top-1 Accuracy from 85.2% to 85.9%, indicating that explicit texture correlation modeling enhances style-sensitive representation. Further incorporating ALGTF increases the accuracy to 86.5%, confirming the effectiveness of adaptively integrating local, global, and texture features. CSHCL also provides consistent improvements by strengthening the discrimination of visually similar style categories. When all three modules are combined, the full model achieves the best performance, reaching 87.1% Top-1 Accuracy and 86.8% Macro F1-score.

4.4. Visualization Analysis

To further evaluate the visual discriminability of different methods, qualitative comparisons are presented on ArtBench-

4.3. Ablation Study

The To investigate the contribution of each proposed component, ablation experiments were conducted on the ArtBench-10 dataset. The original MTH was used as the baseline, and LTCM, ALGTF, and CSHCL were progressively introduced. The results are shown in Table 2.

10 and Painting91_Styles. Two representative samples were selected from each dataset, and the outputs of EfficientNetV2-S, MobileViT-S, Swin-Tiny, ConvNeXt-T, Original MTH, and the proposed method were compared.

As shown in Fig. 2, on the ArtBench-10 dataset, the proposed method preserves clearer structural details and more faithful color transitions than the competing methods. For the first sample, the coastal scene reconstructed by EfficientNetV2-S and MobileViT-S appears noticeably blurred, while Swin-Tiny and ConvNeXt-T provide relatively improved global structure. Original MTH produces a sharper result, but the proposed method further improves fine details and edge consistency, resulting in the best visual fidelity. A similar trend can be observed in the second sample, where the proposed method retains more complete shape boundaries and stronger color contrast in the stylized painting.

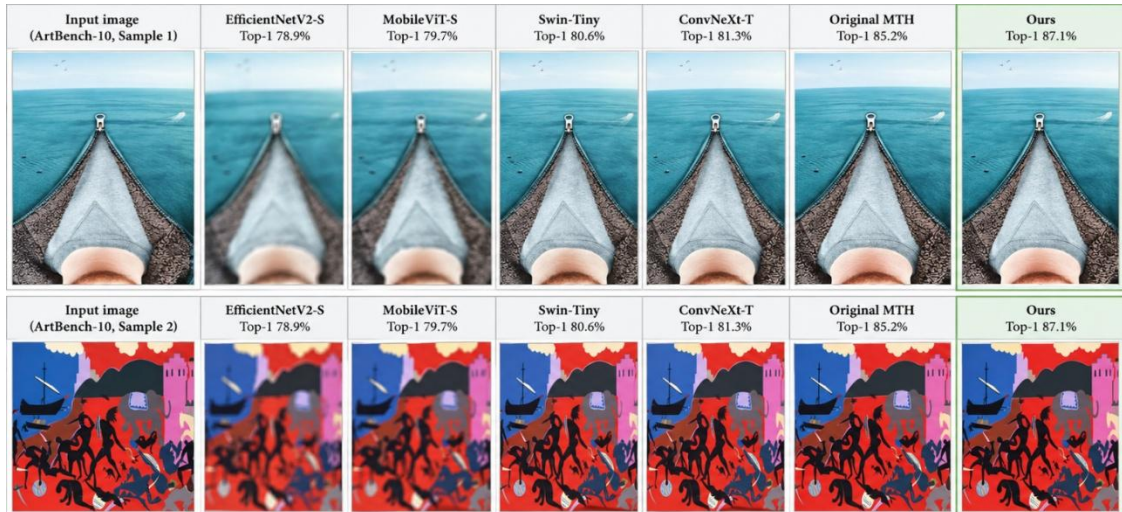


Figure 2. Qualitative comparison on ArtBench-10.

Fig. 3 presents the qualitative comparison on Painting91_Styles, which is more challenging because the artworks contain stronger brushstroke variation and more abstract style characteristics. The proposed method reconstructs the main compositional elements more clearly and preserves more recognizable brushstroke patterns than the other methods. In Sample 1, the baseline methods tend to

smooth local texture details, whereas the proposed model generates a more faithful representation of the original abstract structure. In Sample 2, the figure outline, color blocks, and local artistic texture are more clearly maintained by the proposed method. These observations are consistent with the quantitative improvements reported in the comparison and ablation experiments.

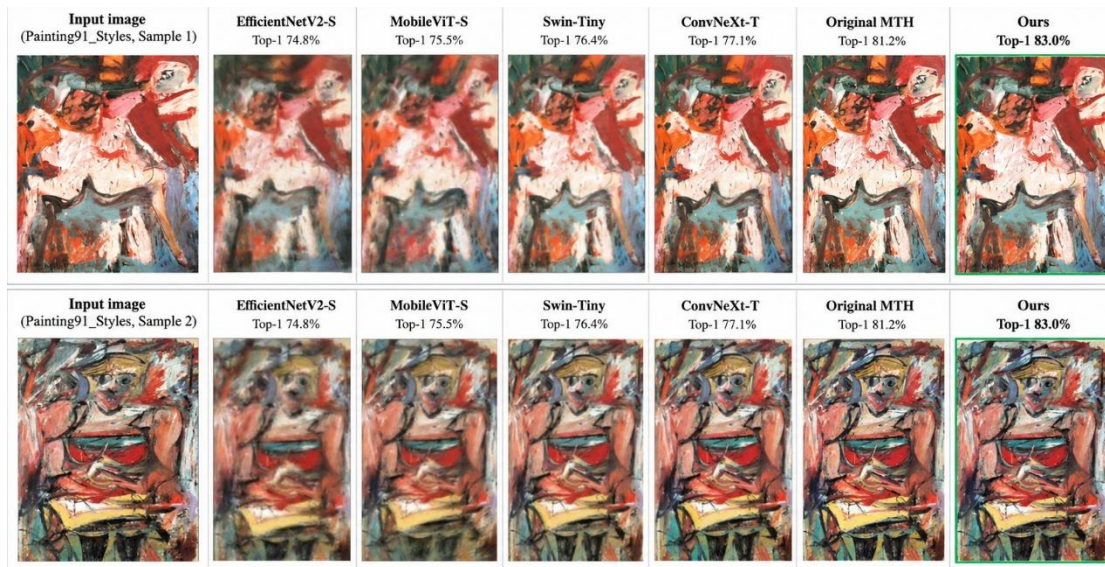


Figure 3. Qualitative comparison on Painting91_Styles.

Overall, the visualization results demonstrate that the proposed framework provides more discriminative artistic style representations by jointly modeling local details, global composition, and texture correlation. The improvements are especially evident for visually complex or abstract artworks, where style-sensitive texture information plays a more important role.

5. Conclusions

This study presents a lightweight artistic style classification framework built on MTH, integrating texture correlation modeling, adaptive feature fusion, and hard contrastive learning. LTCM extracts multi-scale channel correlations from intermediate backbone features to complement conventional local representations. ALGTF dynamically balances local details, global composition, and texture information, while DCSAM refines the fused representation through channel and spatial weighting. During training, CSHCL emphasizes visually similar inter-class samples to encourage more discriminative style representations without introducing an additional contrastive branch during inference. The resulting framework produces style probabilities and category predictions through a unified classification pipeline.

The proposed design targets complementary evaluation settings provided by ArtBench-10 and Painting91_Styles. Its classification gains and computational overhead require confirmation through controlled experiments under consistent data partitions and training protocols. Potential limitations include ambiguous style boundaries, limited training samples, and sensitivity to mislabeled hard negatives. Future work will investigate noise-aware negative selection, more efficient texture aggregation, and evaluation across independent artwork collections. Class-specific attribution analysis will also be explored to examine whether the learned representations emphasize meaningful stylistic characteristics rather than incidental image content.

Acknowledgements

This research was funded by the Key R&D Program of Gansu Province under Grant 24YFFA029, in part by the Key R&D Project of Gansu Provincial Academy of Sciences under Grant 2025PTXM-03, in part by the Youth Fund Project of

Gansu Academy of Sciences under Grant 2025QN-01, and in part by the Achievement Transformation and Technical Application Service Project of Gansu Academy of Sciences 2026CY-02.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflict of interest

There is no conflict of interest.

References

- [1] Ma, K., Lee, S., Ma, X., & Chen, H. (2025). Fine art image classification and design methods integrating lightweight deep learning. *Scientific Reports*, 15, Article 33006. <https://doi.org/10.1038/s41598-025-18420-0>
- [2] Wang, Q., & Feng, G. (2025). Multimodal style aggregation network for art image classification. *Signal Processing: Image Communication*, 137, Article 117309. <https://doi.org/10.1016/j.image.2025.117309>
- [3] Liang, S., Pan, L., Mohammadzadeh, F., et al. (2025). Enhancing artistic style classification through a novel ArtFusionNet framework. *Scientific Reports*, 15, Article 36178. <https://doi.org/10.1038/s41598-025-20087-6>
- [4] Wang, M., Xia, F., Yang, T., Zou, Y., & Zhou, Y. (2026). An art style classification network integrating contrastive learning and counterfactual attention. *PLOS ONE*, 21(6), Article e0343954. <https://doi.org/10.1371/journal.pone.0343954>
- [5] Xiao, B., & Li, H. (2025). Painting style classification and art image analysis model by fusion of convolutional neural network and principal component analysis. *Journal of Computational Methods in Sciences and Engineering*, 25(5), 4048–4060. <https://doi.org/10.1177/14727978251337910>
- [6] Xu, B., & Lu, X. (2026). Artistic style classification in traditional cultural heritage using convolutional neural networks. *Scientific Reports*, 16, Article 20389. <https://doi.org/10.1038/s41598-026-49996-w>
- [7] Zhang, H. (2026). Deep learning approach for art style recognition and colour restoration in visual art images. *Scientific Reports*, 16, Article 24487. <https://doi.org/10.1038/s41598-026-51895-z>

- [8] Li, W. (2025). Enhanced automated art curation using supervised modified CNN for art style classification. *Scientific Reports*, 15, Article 7319. <https://doi.org/10.1038/s41598-025-91671-z>
- [9] El Vaigh, C. B., Garcia, N., Renoust, B., et al. (2025). GNNBoost: Boosting artwork classification with graph embeddings. *Multimedia Tools and Applications*, 84, 39353–39373. <https://doi.org/10.1007/s11042-025-20699-0>
- [10] Elkin, G., Shahr, O. I., Ohayon, Y., Alali, N., & Ben-Shahar, O. (2025). Recognizing artistic style of archaeological image fragments using deep style extrapolation. In *Culture and Computing, HCII 2025 (Lecture Notes in Computer Science, Vol. 15800, pp. 115–131)*. https://doi.org/10.1007/978-3-031-93160-4_8
- [11] Luo, M., Liu, L., Lu, Y., & Suen, C. Y. (2025). Art style classification via self-supervised dual-teacher knowledge distillation. *Applied Soft Computing*, 174, Article 112964. <https://doi.org/10.1016/j.asoc.2025.112964>
- [12] Liu, Y., Bai, H., & Wang, J. (2024). Fine-art recognition using convolutional transformers. *PeerJ Computer Science*, 10, Article e2409. <https://doi.org/10.7717/peerj-cs.2409>
- [13] Lee, S., Choi, J., & Kim, H. J. (2025). EfficientViM: Efficient vision Mamba with hidden state mixer based state space duality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14923–14933).
- [14] He, H., Zhang, J., Cai, Y., et al. (2025). MobileMamba: Lightweight multi-receptive visual Mamba network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [15] Zhang, J., Hu, T., He, H., et al. (2025). EMov2: Pushing 5M vision model frontier. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11), 10560–10576. <https://doi.org/10.1109/TPAMI.2025.3596776>
- [16] Tschannen, M., Gritsenko, A., Wang, X., et al. (2025). SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint* arXiv:2502.14786. <https://doi.org/10.48550/arXiv.2502.14786>
- [17] Siméoni, O., Vo, H. V., Seitzer, M., et al. (2025). DINOv3. *arXiv preprint* arXiv:2508.10104. <https://doi.org/10.48550/arXiv.2508.10104>
- [18] Balmaseda, V., Wang, B., Lin, C.-L., & Yang, T. (2025). Discovering global false negatives on the fly for self-supervised contrastive learning. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267, pp. 2697–2714)*.
- [19] Zhou, Z.-H., Wang, J.-J., Wei, T., & Zhang, M.-L. (2025). Weakly-supervised contrastive learning for imprecise class labels. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267, pp. 79328–79365)*.
- [20] Hieu, N. M., & Ledent, A. (2025). Generalization analysis for supervised contrastive representation learning under non-IID settings. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267, pp. 23179–23218)*.