IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 2, April, 2026, pp. 909-919
Journal homepage: http://iieta.org/journals/ts

Artistic Image Style Recognition Based on Multimodal Visual Features

Zhaoxia Wang

Hebei Art & Design Academy, Baoding 071051, China

Corresponding Author Email: wzx58616@163.com

https://doi.org/10.18280/ts.430228

Received: 3 November 2025
Revised: 25 February 2026
Accepted: 9 March 2026
Available online: 30 April 2026

Keywords:
artistic image style recognition, multimodal visual features, cross-modal attention fusion, discrete wavelet transform, contrastive learning, image processing

ABSTRACT

Artistic image style recognition is a crucial research area at the intersection of computer vision and digital art, with significant applications in digital art archiving and cultural heritage preservation. Existing deep learning approaches often rely on single-modal feature representations, which are insufficient to fully capture style characteristics and lack discriminative power, limiting their accuracy and robustness. To address these challenges, this paper proposes a multimodal visual feature-based method for artistic image style recognition to enhance recognition performance. The method constructs a four-dimensional complementary modality representation, including global semantics, local textures, color distribution, and frequency-domain features. Frequency-domain features are extracted via discrete wavelet transform to enrich the modality completeness. A cross-modal attention fusion mechanism is designed to enable adaptive weighting and deep interaction among modalities. Additionally, a learnable local binary pattern (LBP) combined with a Transformer encoder enhances local texture capture. Contrastive learning is incorporated as a regularization to strengthen the discriminability of style features in the feature space. Experiments on the *WikiArt* and *Painter by Numbers* datasets demonstrate that the proposed method outperforms state-of-the-art approaches in both accuracy and F1-score, showing particularly strong performance in distinguishing visually similar art styles. This study provides a novel technical framework for artistic image style analysis and enriches research and practical approaches in multimodal image processing.

1. INTRODUCTION

Artistic image style recognition, as an important research direction at the intersection of computer vision and digital art [1-3], plays a key role in practical scenarios such as digital art archiving [4], cultural heritage preservation [5], and intelligent art retrieval [6], and also provides new entry points and research paradigms for interdisciplinary research in the field of image processing, possessing significant academic value and application prospects. The expression of artistic style has multidimensional attributes, covering multiple levels including global composition, local brushwork, color tone, and frequency-domain structure [7, 8]. Existing recognition methods are difficult to comprehensively capture these multidimensional features [9, 10], resulting in consistently low recognition accuracy for similar style categories, which cannot meet the high-precision and robustness requirements in practical applications, becoming a core problem urgently needing to be solved in the current field.

Existing artistic image style recognition methods still have many limitations, restricting further improvement of recognition performance. Methods relying on single-modal features cannot cover the complete attributes of artistic styles [11, 12], and are prone to feature redundancy or loss of key information; multimodal fusion methods mostly adopt simple concatenation or weighted summation [13, 14], failing to fully explore the intrinsic correlations between different modalities, and cannot adaptively adjust the weight allocation of each modality according to the style characteristics of the input image; local texture extraction mostly relies on traditional handcrafted features [15, 16], which are difficult to adapt to the complex and variable brushstroke patterns in artistic images, and cannot effectively capture micro-level differences in style; at the same time, insufficient optimization of feature discriminability [17, 18] results in blurred feature boundaries for similar styles, making precise distinction difficult. These limitations collectively constitute the research bottleneck in the current field of artistic image style recognition, and new technical solutions are urgently needed for breakthrough.

To address the above research deficiencies, this paper proposes a method for artistic image style recognition based on multimodal visual features, aiming to improve the accuracy and robustness of style recognition. The main research contributions of this paper are reflected in four aspects: first, constructing a four-dimensional complementary multimodal feature representation, introducing frequency-domain features to improve modality completeness, and achieving comprehensive characterization of the multidimensional attributes of artistic styles; second, designing a cross-modal attention fusion mechanism to achieve deep interaction and adaptive weight learning among modalities, enhancing the discriminative ability of the fused features; third, proposing a local texture enhancement strategy, combining learnable features with a Transformer encoder to strengthen the

capability of capturing micro-level textures such as artistic brushstrokes; fourth, introducing contrastive learning regularization constraints to optimize the feature space distribution and enhance the feature distinguishability of similar styles. Experiments are conducted on two publicly available art datasets, *WikiArt* and *Painter by Numbers*, using accuracy and F1-score as the core evaluation metrics to verify the effectiveness of the proposed method. The structure of the paper is arranged as follows: first, reviewing the related research; second, detailing the core design and technical details of the proposed method; third, verifying the method performance through ablation experiments and comparative experiments; then discussing the advantages and limitations of the method; and finally summarizing the paper and presenting future research directions.

2. Method

2.1 Overall framework of the method

The proposed method for artistic image style recognition adopts an end-to-end deep learning framework. It consists of three core components: multimodal feature extraction branches, cross-modal attention fusion module, and classification optimization module. These modules work collaboratively to realize the complete recognition process from image input to style category output. The overall framework of the method is shown in Figure 1. The input artistic image is first fed into four parallel feature extraction branches, which extract global semantic, local texture, color distribution, and frequency-domain features, providing comprehensive and complementary style representations for the subsequent fusion process. Then, these multimodal features are input into the cross-modal attention fusion module, which deeply explores the intrinsic correlations between modalities and adaptively learns the weights of each modality to generate discriminative fused features. Finally, the fused features are sent into the classification optimization module, mapped by the classification head into the probability distribution of each art style, and constrained by a joint loss function during model training to further optimize feature discriminability and model generalization capability.

2.2 Design of multimodal feature completeness

To comprehensively characterize the multidimensional attributes of artistic styles and overcome the limitations of single-modal and incomplete feature representation in existing methods [19], this paper introduces frequency-domain features based on discrete wavelet transform into artistic image style recognition for the first time, constructing a four-dimensional complementary modality representation consisting of global semantic, local texture, color distribution, and frequency-domain features, achieving a comprehensive description of artistic styles from macro to micro, from spatial to frequency domain. Figure 2 shows the structure of the frequency-domain feature extraction module based on discrete wavelet transform. The core innovation of this design is to supplement spatial features with frequency-domain features that can capture scale structure information which spatial features cannot, addressing the problem that traditional multimodal methods lack the frequency-domain dimension and cannot distinguish intrinsic structural differences in style, forming a multidimensional and non-redundant style feature representation system.

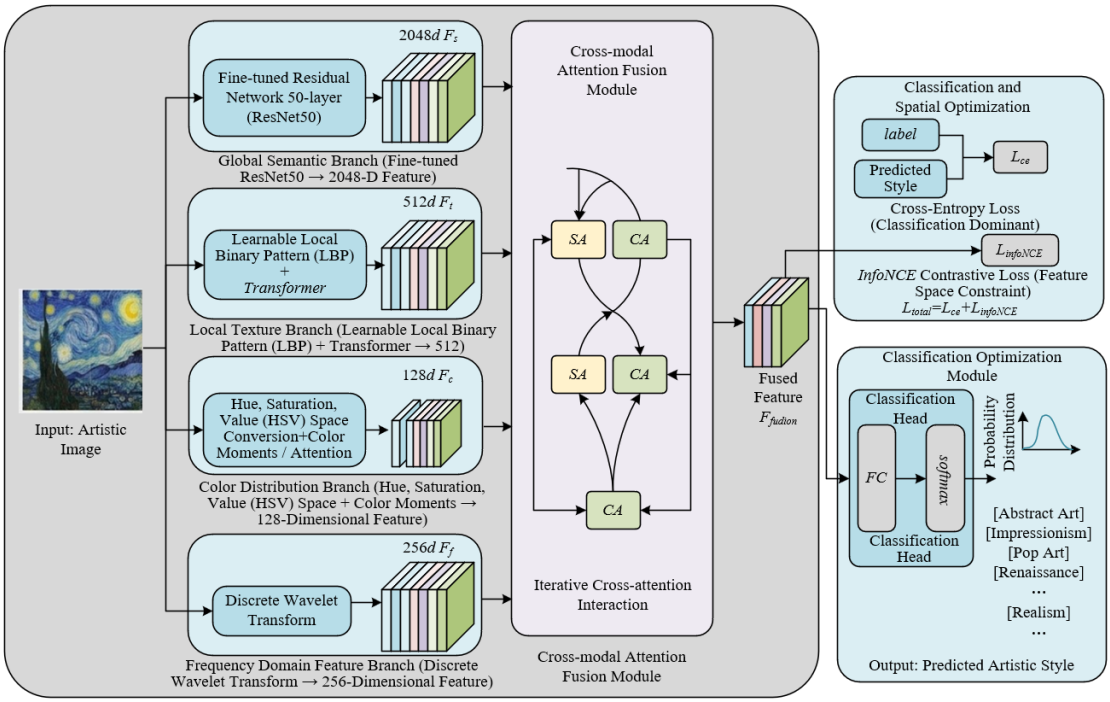


Figure 1. Overall framework of the proposed artistic image style recognition method

The frequency-domain feature branch is the core innovation of this design, and its specific implementation process is as follows: A one-level two-dimensional discrete wavelet transform is applied to the input image using the Haar wavelet basis, decomposing the image into four sub-bands: low-frequency approximation component, horizontal detail component, vertical detail component, and diagonal detail component. Each sub-band has a size of 1/4 of the original

image, corresponding to the overall contour and details in different directions. Each sub-band is independently input into a three-layer convolutional network for feature extraction. The convolution layers all use 3 × 3 kernels, stride 1, and padding 1, with output channels of 64, 128, and 256 respectively. Rectified Linear Unit (ReLU) activation function is applied to introduce nonlinearity, and a batch normalization layer is added after each layer to suppress overfitting. The feature maps of the four sub-bands are concatenated along the channel dimension, transformed into a fixed-length feature vector via global average pooling, and then compressed through a fully connected layer with output dimension of 256, generating the final frequency-domain feature vector. The generation process of the frequency-domain feature vector can be expressed as:

$$f_f=\mathrm{FC}(\mathrm{GAP}(\mathrm{Concat}(C_{LL},C_{LH},C_{HL},C_{HH}))) \quad (1)$$

where, C_{LL}, C_{LH}, C_{HL}, and C_{HH} are the feature maps extracted by the convolutional network from the four sub-bands, *Concat* represents concatenation along the channel dimension, *GAP* is global average pooling, *FC* is fully connected layer, and $f_f \in R^{256}$ is the final output frequency-domain feature vector.

The other three feature branches serve as auxiliary supplements, forming complementary collaboration with the frequency-domain feature to further improve the completeness of multimodal representation. The global semantic branch uses a fine-tuned Residual Network 50-layer (ResNet50) as the backbone network. After removing the top classification layer, a global semantic feature vector of dimension 2048 is output via global average pooling to characterize the overall composition and macroscopic style tendency of the image. The local texture branch combines learnable local binary pattern (LBP) with a Transformer encoder, outputting a local texture feature vector of dimension 512, focusing on capturing micro-level texture details such as artistic brushstrokes. The color feature branch converts the image to Hue, Saturation, Value (HSV) space, combined with color moments and color attention mechanism, outputting a color feature vector of dimension 128 to characterize the color tone preference of artistic styles. The dimensions of the four feature vectors are 2048, 512, 128, and 256 respectively, covering different dimensions of style including macro, micro, color tone, and structure, and achieve comprehensive and accurate characterization of artistic styles through complementary collaboration, providing high-quality feature foundation for subsequent cross-modal fusion.

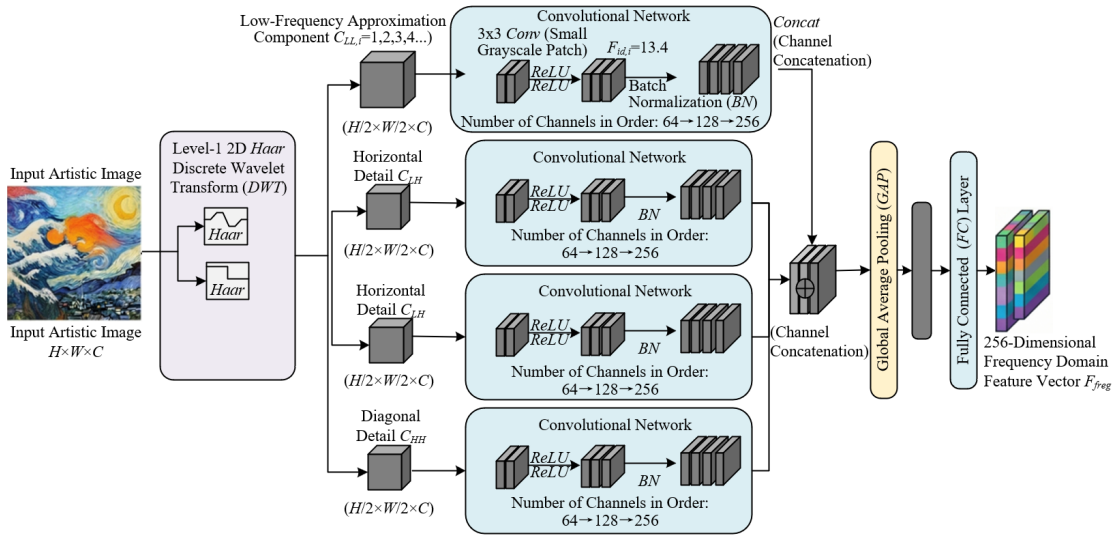


Figure 2. Structure of the frequency-domain feature extraction module based on discrete wavelet transform

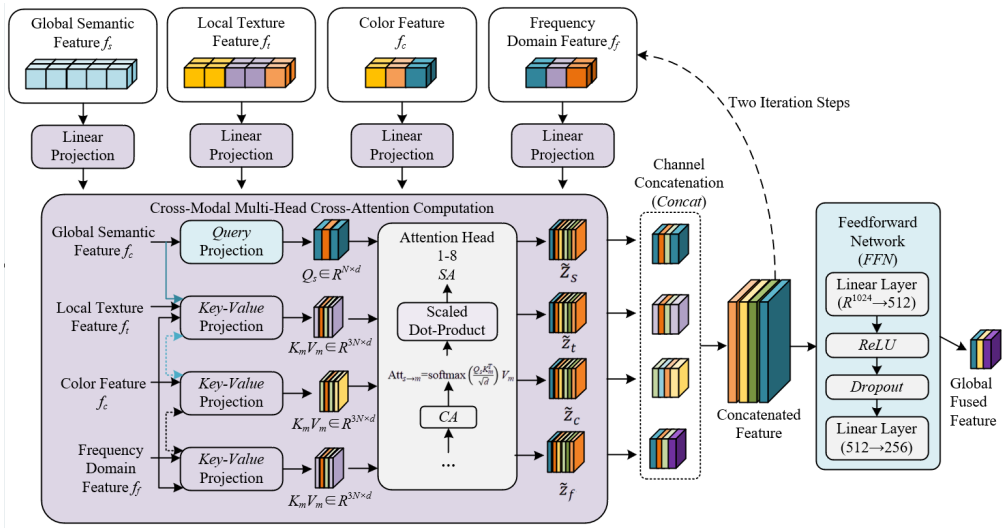


Figure 3. Adaptive fusion mechanism based on multi-head cross-attention

2.3 Cross-modal attention fusion mechanism

To fully explore the intrinsic correlations between multimodal features and to address the limitations of existing fusion methods with fixed weights and insufficient modality interaction [20], and to achieve efficient complementarity and adaptive fusion of multimodal features, this paper proposes an adaptive fusion module based on multi-head cross-attention. The schematic diagram of the module is shown in Figure 3. The core innovation of this module is that, with multi-head cross-attention as the core, it dynamically learns the importance weights of each modality feature, strengthens the contribution of key modality information, suppresses redundant information interference, and deepens the interaction degree between modalities through multiple iterations, generating discriminative fused features that provide high-quality feature support for subsequent style classification.

Feature projection is the basic prerequisite for implementing cross-modal fusion. Its purpose is to map multimodal features of different dimensions into the same feature space, eliminating fusion bias caused by dimensional differences. Let the original feature vectors of the four modalities, global semantic, local texture, color, and frequency-domain, be $f_s \in R^{2048}$, $f_t \in R^{512}$, $f_c \in R^{128}$, and $f_f \in R^{256}$, respectively. Through independent linear transformations, they are mapped to the target dimension d=256, obtaining the projected features $z_s, z_t, z_c, z_f \in R^{256}$. The linear transformation is calculated as:

$$z_m = f_m W_m + b_m \tag{2}$$

where, $m \in \{s, t, c, f\}$ corresponds to the four modalities, W_m is the linear projection matrix with dimensions 2048 × 256, 512 × 256, 128 × 256, 256 × 256, respectively, $b_m \in R^{256}$ is the bias term, adaptively learned through model training. The projected features preserve the core style information of each modality and provide a unified feature basis for subsequent cross-attention computation.

Multi-head cross-attention calculation is the core innovation of this module. By constructing a "query-key-value" interaction pattern, it achieves deep correlation mining between modalities and adaptive weight allocation. Each modality's projected feature is used as a query, and the projected features of the other three modalities are used as keys and values simultaneously, constructing four cross-attention structures. Each attention structure uses an 8-head multi-head attention mechanism, with each head having a feature dimension of 256/8 = 32. Taking the global semantic feature as the query as an example, its attention computation process is as follows: first, the query, key, and value are mapped to the same dimension through learnable matrices, i.e., $Q_s = z_s W_Q$, $K_m = z_m W_K$, $V_m = z_m W_V$, where $W_Q, W_K, W_V \in R^{256\times256}$ are learnable parameters; then, the similarity between the query and each key is computed and normalized through softmax to obtain attention weights, and finally the attention output is calculated by combining the value:

$$\mathrm{Att}_{s\to m} = \mathrm{softmax}\left(\frac{Q_s K_m^T}{\sqrt{d}}\right) V_m \tag{3}$$

where, d is the scaling factor, used to alleviate bias in similarity computation caused by increased dimensionality. The attention outputs corresponding to the three modalities are concatenated, passed through a linear transformation and ReLU activation function, and the fused feature dominated by global semantic features $\tilde{z}_s$ is obtained. Using the same process, local texture, color, and frequency-domain features are used as queries to obtain the corresponding fused features $\tilde{z}_t, \tilde{z}_c, \tilde{z}_f$.

To further deepen the interaction between modalities and enhance the discriminability of fused features, the above cross-attention computation process is performed twice iteratively, updating the fused features of each modality in each iteration to strengthen complementary information between modalities. After iteration, the four modality-dominated fused features $\tilde{z}_s, \tilde{z}_t, \tilde{z}_c, \tilde{z}_f$ are concatenated along the channel dimension, forming a feature vector of dimension 4 × 256 = 1024, which is then input into a feedforward network for feature optimization and dimensionality reduction. The feedforward network consists of two fully connected layers, with a hidden layer dimension of 512, using ReLU activation and a dropout layer (dropout rate 0.5) to suppress overfitting. The output layer reduces the feature dimension to 256, obtaining the final multimodal fused feature $F \in R^{256}$, calculated as:

$$F = FC_2(ReLU(FC_1(Concat(\tilde{z}_s, \tilde{z}_t, \tilde{z}_c, \tilde{z}_f)))) \tag{4}$$

Compared with traditional simple concatenation or weighted summation fusion methods, the proposed cross-modal attention fusion mechanism has significant advantages: it can dynamically learn the importance weights of each modality, adaptively adjust the contribution of different modalities according to the style characteristics of the input image, and capture the intrinsic correlations between modalities through cross-attention, effectively suppressing redundant information and highlighting feature cues that play a key role in style recognition, thereby significantly enhancing the discriminability of fused features and providing a foundation for improving subsequent style classification performance.

2.4 Local texture enhancement strategy

The micro-level textures such as brushstrokes and surface texture in artistic images are the core cues for distinguishing different art styles. Traditional local texture extraction methods are difficult to adaptively capture the complex and variable brushstroke patterns in artistic images [21], and cannot effectively mine long-range dependencies between local patches. To address this problem, this paper proposes a local texture enhancement strategy. The core innovation lies in combining learnable LBP with a Transformer encoder to achieve precise capture and deep representation of micro-level textures in artistic images, improving the network's sensitivity to style details and providing high-quality local texture support for multimodal feature fusion. Figure 4 shows the local texture enhancement strategy and feature extraction process.

The learnable LBP is the core improvement of this strategy, aiming to overcome the limitations of traditional LBP being non-differentiable, unable to train end-to-end, and having limited feature representation capability, enabling adaptive learning of artistic brushstroke textures. In this paper, the traditional LBP is extended to a differentiable continuous form, using convolutional layers to learn the weight distribution of neighboring pixels instead of using fixed binary coding based on grayscale comparison. Specifically, a 3 × 3 convolution kernel corresponding to an 8-neighborhood structure is used,

with the convolution kernel parameters as learnable weights, convolving the local region of the input image to generate a continuously valued texture response map. Its computation process can be expressed as:

$$T(x,y)=\sum_{i=0}^{7} w_i \cdot I(x+i_x,y+i_y) \quad (5)$$

where, $T(x,y)$ is the response value at coordinate (x,y) in the texture response map, w_i is the learnable neighborhood weight, $I(x+i_x,y+i_y)$ is the grayscale value of the corresponding pixel in the neighborhood, and i_x, i_y are the coordinate offsets of the neighboring pixel relative to the center pixel. This design adaptively adjusts neighborhood weights through model training, accurately capturing unique brushstroke patterns of different art styles. Compared with traditional LBP with fixed coding, it is more flexible and can effectively adapt to complex textures such as Van Gogh's swirling brushstrokes and Monet's pointillism, and its differentiable property allows integration into end-to-end deep learning frameworks for collaborative optimization with subsequent modules.

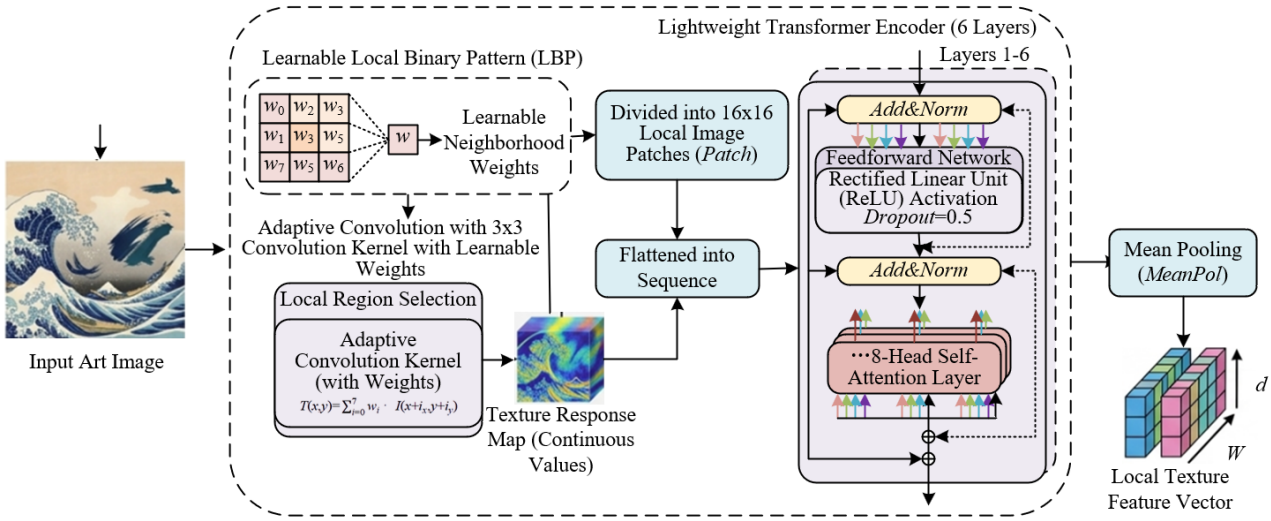


Figure 4. Local texture enhancement strategy and feature extraction process

To further explore long-range dependencies between local texture patches and strengthen the global correlation of texture features, after generating the texture response map by learnable LBP, a lightweight Transformer encoder is introduced for feature enhancement. The texture response map is first divided into 16 × 16 local image patches, each patch is flattened into a one-dimensional feature vector to form the input sequence. The encoder has six layers, each consisting of a multi-head self-attention mechanism and a feedforward network. The multi-head self-attention is designed with 8 heads, each head having a feature dimension of 64, hidden layer dimension of 256, and the feedforward network uses ReLU activation. A dropout layer with dropout rate 0.5 is added in each layer to suppress overfitting. The Transformer encoder computes the correlation weights between different local patches through self-attention, capturing cross-region texture dependencies. The output sequence of the encoder is transformed into a fixed-length local texture feature vector $f_t \in R^{512}$ through mean pooling, expressed as:

$$f_t=\text{MeanPool}(\text{Transformer}(\text{Patch}(T(x,y)))) \quad (6)$$

where, *Patch* represents the image patch division and flattening operation, *Transformer* is the 6-layer lightweight encoder, and *MeanPool* is the mean pooling operation.

This local texture enhancement strategy has good adaptability to artistic images and can accurately capture micro-level texture features of different art styles: for Van Gogh's swirling brushstrokes, the learnable LBP strengthens the response of swirling textures by adjusting neighborhood weights, and the Transformer encoder captures cross-region correlations of the swirling textures; for Impressionist pointillism brushstrokes, this strategy effectively distinguishes the density and distribution characteristics of pointillist strokes, compensating for the insufficient capture of micro-level details by global semantic and color modalities. The local texture feature output of this branch forms a complementary collaboration with global semantic, color, and frequency-domain features, capable of depicting micro-level brushstroke details and integrating texture information globally through the Transformer encoder, further improving the completeness of multimodal features and providing critical texture feature support for subsequent cross-modal fusion and style classification.

2.5 Contrastive learning regularization optimization

The discriminability of fused features directly determines the accuracy of artistic image style recognition, especially in similar style recognition tasks. Existing methods generally have problems of feature boundary ambiguity and insufficient discriminative capability [22, 23]. To address this problem, this paper proposes a contrastive learning regularization optimization strategy. The core innovation lies in introducing Information Noise-Contrastive Estimation (InfoNCE) contrastive loss jointly with cross-entropy loss for optimization. The main loss supervises category classification, and the regularization term constrains the feature space distribution, which together enhance the discriminability of fused features, effectively solving the core difficulty of distinguishing similar style features and providing assurance for high-precision style recognition.

The reasonable design of the loss function is the key to optimization effectiveness. The total loss function constructed

in this paper is a weighted sum of cross-entropy loss and InfoNCE contrastive loss, expressed as:

$$L=L_{CE}+\lambda L_{CL} \tag{7}$$

where, L_{CE} is the cross-entropy loss, serving as the main loss to supervise the model's category classification task; L_{CL} is the InfoNCE contrastive loss, serving as the regularization term to constrain the spatial distribution of fused features; $\lambda = 0.1$ is the balancing coefficient, set based on multiple experimental validations. This coefficient can effectively balance the contributions of the main loss and the regularization term, avoiding underfitting caused by excessive regularization or insufficient constraint effect, ensuring the model simultaneously considers classification accuracy and feature discriminability. The specific calculation formula of the cross-entropy loss is:

$$L_{CE}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \tag{8}$$

where, N is the training batch size, C is the number of art style categories, $y_{i,c}$ is the *one-hot* encoding of the ground truth label of the i-th sample for the c-th style category, and $\hat{y}_{i,c}$ is the predicted score of the i-th sample belonging to the c-th style category. This loss supervises the model to learn the mapping between fused features and style categories by measuring the difference between the predicted distribution and the true distribution, ensuring the model can accurately distinguish different style categories, and serves as the core supervisory signal for style classification.

InfoNCE contrastive loss is the core innovative part of this optimization strategy. Its core function is to constrain the spatial distribution of fused features, strengthening the aggregation of same-style features and the separation of different-style features, especially improving the feature distinction ability for similar styles. The specific calculation formula is:

$$L_{CL}=-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathrm{sim}(F_i,F_{i+})/\tau)}{\sum_{j=1}^{N}1_{|j\neq i|}\exp(\mathrm{sim}(F_i,F_j)/\tau)} \tag{9}$$

where, F_i is the multimodal fused feature of the i-th sample, F_{i+} is the positive sample feature belonging to the same style as F_i, $\mathrm{sim}(,)$ denotes cosine similarity used to measure the similarity between two feature vectors, and $\tau = 0.1$ is the temperature coefficient, set to alleviate gradient vanishing in similarity calculation and make the feature distribution more distinguishable. Positive and negative sample pairs are constructed within the training batch. Positive pairs are selected from different samples of the same style, and negative pairs are selected from samples of different styles. Under the constraint of this loss function, fused features of the same style are forced to aggregate in the embedding space, and fused features of different styles are pushed apart, effectively reducing intra-style feature differences and enlarging inter-style feature distances, especially solving the problem of ambiguous feature boundaries for similar styles.

The classification head serves as the bridge between fused features and style categories, and its structural design directly affects classification performance and model generalization ability. The classification head designed in this paper consists of two fully connected layers. The input is the multimodal fused feature $F\in R^{256}$. The first fully connected layer maps the input feature to 512 dimensions. A ReLU activation function is applied to introduce nonlinearity and enhance the feature fitting ability of the model, and a dropout layer with a dropout rate of 0.5 is added to suppress overfitting, preventing the model from over-relying on noise in training data. The second fully connected layer maps the 512-dimensional feature to the number of style categories C, outputting predicted scores $\hat{y}$ for each style category, serving as the basis for cross-entropy loss calculation. This classification head structure is simple and efficient. Through the synergistic effect of the dropout layer and ReLU activation, it ensures effective mapping of features while preventing overfitting. In combination with the joint loss function optimization, it further improves the accuracy and robustness of style recognition.

3. EXPERIMENTS AND RESULT ANALYSIS

To verify the effectiveness and superiority of the proposed artistic image style recognition method based on multimodal visual features, systematic experiments were designed, including ablation experiments, comparative experiments, and stability and generalization tests. The experiments use the *WikiArt* and *Painter by Numbers* public art datasets as carriers, and through precise experimental settings, scientific metric evaluations, and in-depth result analysis, comprehensively verify the performance advantages of the proposed method and the core contributions of each innovation.

3.1 Experimental settings

The experiments adopt two internationally recognized public datasets for artistic image style recognition to ensure the objectivity and comparability of the experimental results. The *WikiArt* dataset contains 81,449 images from 195 artists, covering 27 classic art styles. The image resolution ranges from 256 × 256 to 1024 × 1024, and the data distribution is relatively balanced, with 1,500 to 5,000 samples per style. The *Painter by Numbers* dataset contains 102,231 images, covering 34 art styles, with a unified image resolution of 256 × 256. Some styles have fewer samples, which further tests the generalization ability of the model. The data preprocessing steps are as follows: all images are normalized to a size of 224 × 224, and RGB channel standardization is applied to accelerate model convergence. To alleviate overfitting, data augmentation strategies are adopted during training, including random horizontal flipping, ±15° random rotation, and 0.8–1.2× brightness adjustment. No data augmentation is applied during testing to ensure the authenticity of test results. The datasets are divided into training, validation, and test sets with a ratio of 8:1:1 for model training, parameter tuning, and performance evaluation.

The experimental hardware environment consists of an NVIDIA RTX 3090 GPU (24GB memory), Intel Core i9-12900K CPU, and 64GB DDR5 memory. The software environment is based on the PyTorch 1.10 deep learning framework, Python 3.8 programming language, and CUDA 11.3 for accelerated computation. The model training parameters are set as follows: Adam optimizer is used with an initial learning rate of $1e$-4, learning rate decay follows a cosine annealing schedule with a decay period of 20 epochs; the total training iterations are 100 epochs, and the batch size

is set to 32; weight decay coefficient is set to 1e-5 to suppress overfitting; the balance coefficient λ between cross-entropy loss and contrastive learning loss is set to 0.1, and the temperature coefficient τ is set to 0.1. All parameter values are determined through multiple controlled experiments to obtain the optimal settings.

3.2 Ablation experiments

The ablation experiments aim to verify the effectiveness of the four core innovations proposed in this paper one by one. Four ablation models were constructed, and their performance was compared with the complete method of this paper to quantify the contribution of each innovation. The ablation experiments were conducted simultaneously on the two datasets, and the results are shown in Table 1.

From the results in Table 1, it can be observed that removing any innovation leads to varying degrees of performance degradation, verifying the necessity and core contribution of each innovation. Removing the frequency domain feature branch results in a 3.44% drop in accuracy on the *WikiArt* dataset and a 3.86% drop on the *Painter by Numbers* dataset, indicating that frequency domain features can supplement scale structure information that spatial features cannot capture, effectively improving the completeness of multimodal features and enhancing the model's ability to distinguish intrinsic style structural differences. Replacing cross-modal attention fusion with simple concatenation reduces accuracy by 4.11% and 4.62% on the two datasets, highlighting the advantage of the cross-modal attention mechanism in exploring inter-modal associations and adaptively assigning weights, which achieves more efficient multimodal feature complementarity than simple concatenation. Replacing learnable LBP+Transformer with traditional LBP reduces accuracy by 2.89% and 3.29%, showing that the combination of learnable LBP and Transformer can more accurately capture micro-textures such as artistic brushstrokes, improving the model's sensitivity to style details. Removing the contrastive learning regularization results in accuracy drops of 2.52% and 2.23%, indicating that contrastive learning effectively constrains the feature space distribution, strengthens the discriminability of similar style features, and further improves model discriminative performance. The collaborative effect of the four innovations allows the proposed method to achieve optimal performance, fully verifying the rationality and effectiveness of the method design.

Table 1. Comparison of ablation experiment results (%)

Model Configuration	*WikiArt*			*Painter by Numbers*		
	Accuracy	Macro F1	Micro F1	Accuracy	Macro F1	Micro F1
Proposed Method	92.76	91.85	92.63	89.54	88.72	89.41
Ablation 1: Remove frequency domain feature branch	89.32	88.17	89.15	85.68	84.53	85.59
Ablation 2: Replace cross-modal attention fusion with simple concatenation	88.65	87.52	88.48	84.92	83.87	84.85
Ablation 3: Replace learnable Local Binary Pattern (LBP)+Transformer with traditional LBP	89.87	88.93	89.74	86.25	85.19	86.18
Ablation 4: Remove contrastive learning regularization	90.24	89.36	90.11	87.31	86.45	87.22

3.3 Comparative experiment results and analysis

The comparative experiments aim to verify the performance differences between the proposed method and existing mainstream methods, highlighting the superiority of the proposed method. The experimental results are shown in Table 2. Additionally, visualization analysis further validates the effectiveness of each module and intuitively demonstrates the advantages of the proposed method.

Table 2. Comparison of experimental results (%)

Method	*WikiArt*			*Painter by Numbers*			Similar Style Recognition Accuracy
	Accuracy	Macro F1	Micro F1	Accuracy	Macro F1	Micro F1	(Abstract Expressionism / Action Painting)
Handcrafted Features +Support Vector Machine (SVM)	68.35	66.72	68.14	65.42	63.89	65.27	52.18
Visual Geometry Group 16-layer Network (VGG16)	79.58	78.34	79.41	76.85	75.62	76.71	64.35
Residual Network (50-layer)	85.27	84.19	85.12	82.63	81.57	82.51	72.46
Multi-modal Representation Learning	89.84	88.97	89.72	86.79	85.83	86.68	78.59
Proposed Method	92.76	91.85	92.63	89.54	88.72	89.41	87.32

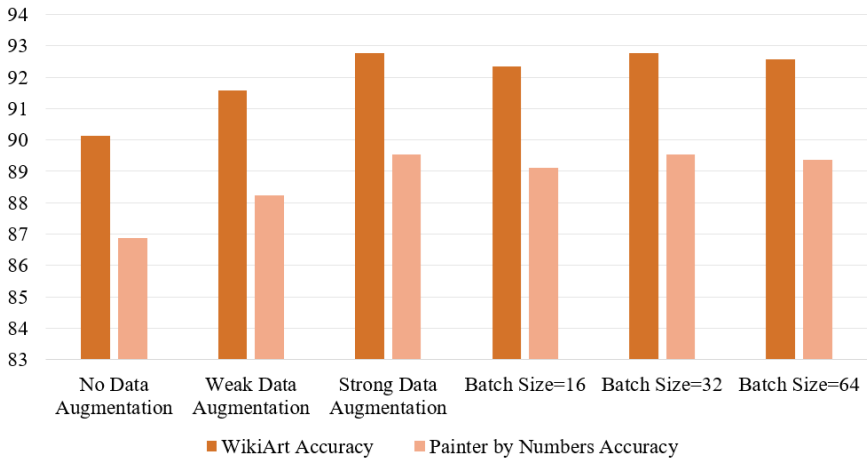
From the results in Table 2, the proposed method significantly outperforms all comparative methods on all evaluation metrics on both datasets, demonstrating strong artistic image style recognition capability. Compared with the traditional method Handcrafted Features + Support Vector Machine (SVM), the proposed method improves accuracy on the *WikiArt* dataset by 24.41% and similar style recognition accuracy by 35.14%, indicating that deep learning-based multimodal feature representation has stronger expressive ability than traditional handcrafted features. Compared with mainstream deep learning methods Visual Geometry Group 16-layer Network (VGG16) and ResNet50, the proposed method increases accuracy by 13.18% and 7.49%, respectively, because it integrates multimodal feature fusion and regularization optimization, addressing the limitations of single-modal incomplete feature representation and insufficient discriminability. Compared with the recent multimodal method Multi-modal Representation Learning (MURAL), the proposed method improves accuracy by 2.92% and 2.75%, and similar style recognition accuracy by 8.73%.

The core advantage lies in the introduction of frequency domain features to improve modal completeness, achieving more efficient inter-modal interaction through cross-modal attention fusion, and further enhancing feature discriminability through contrastive learning, effectively solving the problem of insufficient accuracy in similar style recognition.

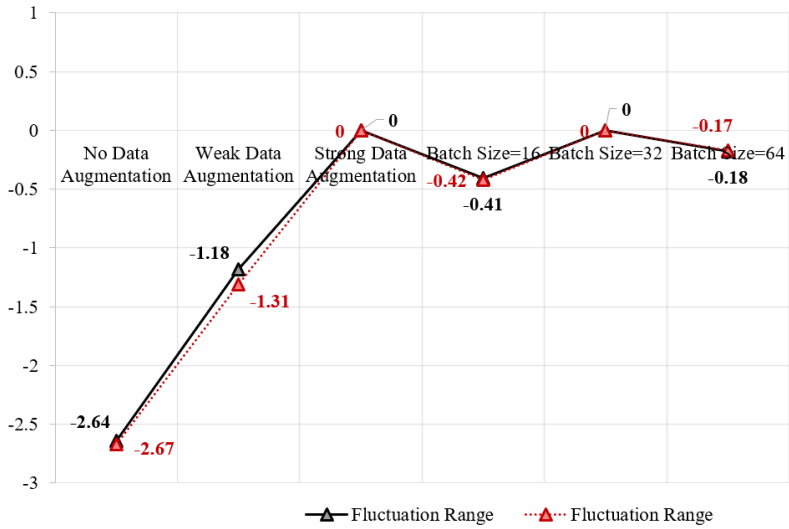
In the similar style recognition task, Abstract Expressionism and Action Painting are difficult to distinguish due to high similarity in brushstrokes and composition. The proposed method achieves an accuracy of 87.32% on this task, significantly higher than other comparative methods, fully demonstrating that the proposed method can effectively capture subtle differences between the two styles and validating the effectiveness of the proposed innovations.

3.4 Stability and generalization tests

To verify the practicality of the proposed method, stability and generalization tests were designed to test the model's performance fluctuations under different experimental conditions and its adaptability across datasets. The experimental results are shown in Figure 5 and Table 3.



(a) Performance situation



(b) Fluctuation situation

Figure 5. Stability test results (%)

From the stability test results in Figure 5, it can be observed that the performance fluctuation of the proposed method under different data augmentation intensities and batch sizes is small, with a maximum fluctuation not exceeding 2.67%. The model achieves optimal performance under strong data augmentation conditions; without data augmentation, performance slightly decreases but still maintains a high level. When the batch size varies within the range of 16–64, the performance fluctuation does not exceed 0.42%, indicating that the proposed method has strong adaptability to data augmentation intensity and batch size, demonstrating good stability and the ability to adapt to different experimental environments and data conditions.

Table 3. Genralization test results (%)

Training Dataset	Testing Dataset	Accuracy	Macro F1	Micro F1
WikiArt	*Painter by Numbers*	83.27	82.15	83.14
Painter by Numbers	*WikiArt*	85.69	84.57	85.58

The generalization test adopts a cross-dataset evaluation, where a model trained on one dataset is tested on another dataset that did not participate in training. From the results in Table 3, the model trained on *WikiArt* achieves an accuracy of 83.27% on the *Painter by Numbers* test set, and the model

trained on *Painter by Numbers* achieves an accuracy of 85.69% on the *WikiArt* test set, both maintaining high recognition accuracy. This indicates that the proposed method can effectively adapt to art image datasets with different distributions, demonstrating good generalization ability. The core reason is that the multimodal features constructed in this paper can comprehensively characterize the intrinsic attributes of artistic styles, while cross-modal attention fusion and contrastive learning optimization enhance the model's adaptability to different styles and datasets, laying the foundation for practical application of the method.

To intuitively verify the effectiveness of the multimodal feature extraction branches in capturing the multidimensional attributes of artistic styles, a feature response visualization experiment was conducted. In Figure 6, Post-Impressionist works by Van Gogh are used as input samples. Subfigure (b) shows the local texture feature response map, which, through the collaborative effect of learnable LBP and the Transformer encoder, precisely enhances micro-texture regions such as brushstrokes and surface textures, verifying the high sensitivity of the local texture enhancement strategy to microscopic style cues and effectively capturing the unique brushstroke characteristics of Post-Impressionism. Subfigure (c) shows the frequency domain feature response map, generated based on discrete wavelet transform, effectively highlighting structural contours and edge details of the painting, supplementing scale structure information that spatial features cannot capture, verifying the role of frequency domain features in completing multimodal representation. Subfigure (d) shows the color feature response map, where the color attention module adaptively highlights core regions conforming to Post-Impressionist color tone preferences, enhancing the discriminability of style color features.

Figure 6. Visualization of multimodal visual feature extraction branch responses

In summary, based on the results of the ablation experiments, comparative experiments, and stability and generalization tests, the proposed multimodal visual feature-based art image style recognition method demonstrates significant performance advantages. The four core innovations work collaboratively to effectively address the problems of single-modal features, inefficient fusion, insufficient texture capture, and poor feature discriminability in existing methods. The proposed method achieves recognition performance superior to current mainstream methods on both public datasets, especially in similar style recognition tasks. Meanwhile, the method demonstrates good stability and generalization, being able to adapt to different experimental conditions and dataset distributions, providing an efficient and robust technical solution for art image style recognition with significant practical and academic value.

4. DISCUSSION

The proposed multimodal visual feature-based art image style recognition method has its core advantages in the synergistic effect of four innovations, effectively overcoming the key bottlenecks of existing methods and showing significant advancement in both academic innovation and performance. The integrity design of multimodal features, by introducing frequency domain features, compensates for the limitation of traditional methods that only focus on spatial features, achieving comprehensive characterization of the multidimensional attributes of artistic styles. The cross-modal attention fusion mechanism abandons the inefficient simple concatenation fusion method, and through dynamic weight allocation and deep modal interaction, maximizes the complementary value of each modality. The local texture enhancement strategy addresses the insufficient adaptability problem of traditional texture extraction methods, accurately capturing micro details such as artistic brushstrokes. The contrastive learning regularization optimization effectively strengthens feature discriminability, particularly enhancing the differentiation of similar styles. Experimental results show that the four innovations do not exist in isolation but form an organically complementary system, collectively solving the core deficiencies of existing methods such as single-modality features, low fusion efficiency, inaccurate texture capture, and blurred feature boundaries. Compared with existing similar multimodal studies, the core difference of this method is the first systematic introduction of frequency domain features into art style recognition, constructing a four-dimensional complementary modal system, and at the same time, achieving dual enhancement of feature representation and discriminability through the deep combination of cross-attention fusion and contrastive learning. Its performance advantage is particularly significant in similar style recognition tasks, enriching the research ideas and technical pathways of multimodal fusion in the field of image processing.

Although the proposed method achieves good recognition performance, there are still certain limitations, reflecting the objectivity and rigor of the study. In frequency domain feature extraction, only single-level discrete wavelet transform is used for decomposition, which can only capture single-scale frequency information, making it difficult to fully mine the structural differences of artistic styles at different scales, and leaving room for improvement in depicting details of complex styles. In terms of model efficiency, the 6-layer Transformer encoder used in the local texture enhancement strategy can effectively capture dependencies between local blocks, but its computational complexity is relatively high, affecting the inference speed and making it difficult to meet the requirements of real-time recognition scenarios. Regarding sample adaptability, model training depends on sufficient sample data, and for minority art styles with fewer samples,

feature learning is insufficient, resulting in relatively lower recognition accuracy and leaving room for improvement in generalization ability. In addition, the robustness of the model to image noise and resolution differences has not been systematically tested, and its adaptability in complex real-world scenarios requires further verification.

In response to the limitations of the proposed method and in combination with research hotspots in the field of image processing, the following specific future research directions are proposed to provide clear extension ideas for subsequent studies. To address the scale limitation of frequency domain feature extraction, multi-scale discrete wavelet transform or wavelet packet transform can be introduced to mine frequency information at different scales and bands, further improving the integrity of multimodal features and enhancing the ability to depict complex artistic styles. To address the problem of high computational complexity, lightweight Transformer structures can be explored, through pruning, knowledge distillation, or using hybrid Convolutional Neural Network (CNN)-Transformer structures, reducing model parameters and computation while ensuring recognition performance, improving inference speed and adapting to real-time application scenarios. To address the insufficient adaptability to minority style samples, *Few-shot* learning, meta-learning, or similar techniques can be introduced, allowing efficient learning from few samples to enhance the model's recognition ability for minority styles and expand the applicability of the method. In addition, future research can further study the robustness of the model to noise and resolution differences, optimize the feature extraction process of each feature branch using attention mechanisms, and explore the integration of multimodal features with prior knowledge in the art domain, further enhancing the academic value and practical applicability of the method.

5. CONCLUSION

This paper addresses the core issues of existing art image style recognition methods, including single-modality features, low fusion efficiency, inaccurate texture capture, and insufficient feature discriminability, and proposes a multimodal visual feature-based art image style recognition method, systematically completing full-process innovations in multimodal feature design, cross-modal fusion, texture enhancement, and loss optimization. The core contributions of this method are reflected in four aspects: constructing a four-dimensional complementary modal representation of global semantics, local textures, color distribution, and frequency domain features, improving modal integrity through the introduction of discrete wavelet transform; designing an adaptive fusion mechanism based on multi-head cross-attention to achieve deep interaction of multimodal features and adaptive weight allocation; proposing a texture enhancement strategy combining learnable LBP and Transformer, improving the capability to capture micro details such as artistic brushstrokes; introducing joint optimization of *InfoNCE* contrastive loss and cross-entropy loss, enhancing the discriminability of fused features and the ability to distinguish similar styles.

Experimental results show that the proposed method significantly outperforms existing traditional methods, mainstream deep learning methods, and recent multimodal methods in accuracy, macro F1 score, and micro F1 score on the *WikiArt* and *Painter by Numbers* public datasets. It performs particularly well in similar style recognition tasks such as Abstract Expressionism and Action Painting, while also demonstrating good stability and generalization, fully achieving the expected research objectives. This study provides an efficient and robust technical framework for art image style recognition, effectively enriching the research ideas and practical approaches of multimodal fusion in the field of image processing, filling the gap of frequency domain features in art style recognition applications, and providing important academic reference and technical support for related research at the intersection of computer vision and digital art.

REFERENCES

[1] Fu, X. (2022). Digital image art style transfer algorithm based on CycleGAN. Computational Intelligence and Neuroscience, 2022: 1-10. https://doi.org/10.1155/2022/6075398

[2] Cai, W. (2022). Chinese painting and calligraphy image recognition technology based on pseudo linear directional diffusion equation. Applied Mathematics and Nonlinear Sciences, 8(1): 1509-1518. https://doi.org/10.2478/amns.2022.2.0139

[3] Wang, Q., Feng, G. (2025). Multimodal style aggregation network for art image classification. Signal Processing: Image Communication, 137: 117309. https://doi.org/10.1016/j.image.2025.117309

[4] Nimis, E. (2014). In search of African history: The re-appropriation of photographic archives by contemporary visual artists. Social Dynamics, 40(3): 556-566. https://doi.org/10.1080/02533952.2014.992598

[5] Yi, J., Tian, Y., Zhao, Y. (2024). Novel approach to protect red revolutionary heritage based on artificial intelligence algorithm and image-processing technology. Buildings, 14(9): 3011. https://doi.org/10.3390/buildings14093011

[6] Wan, J.X., Yu, X.B. (2021). Intelligent retrieval method of approximate painting in digital art field. Scientific Programming, 2021: 5796600. https://doi.org/10.1155/2021/5796600

[7] Zhao, H., Zhang, B., Yang, Y. (2025). Contrastive attention and fine-grained feature fusion for artistic style transfer. Journal of Visual Communication and Image Representation, 110. https://doi.org/10.1016/j.jvcir.2025.104451

[8] Li, W., Cheng, S. (2026). Style and structure-aware generative super-resolution for bashu painting and calligraphy restoration. International Journal of Knowledge Management, 22(1): 1-21. https://doi.org/10.4018/ijkm.398364

[9] Tan, Y. (2022). Feature recognition and style transfer of painting image using lightweight deep learning. Computational Intelligence and Neuroscience, 2022: 1-10. https://doi.org/10.1155/2022/1478371

[10] Ahmadkhani, S., Moghaddam, M.E. (2022). Image recommender system based on compact convolutional transformer image style recognition. Journal of Electronic Imaging, 31(04): 043054. https://doi.org/10.1117/1.jei.31.4.043054

[11] Huang, D. (2018). Intelligent recognition method of micro image feature recognition in large data

environment. Journal of Coastal Research, 83(sp1): 697-705. https://doi.org/10.2112/SI83-116.1

[12] Cheng, Y., Liu, K., Yang, J. (1993). A novel feature extraction method for image recognition based on similar discriminant function (SDF). Pattern Recognition, 26(1): 115-125. https://doi.org/10.1016/0031-3203(93)90093-c

[13] Shen, D., Zareapoor, M., Yang, J. (2021). Multimodal image fusion based on point-wise mutual information. Image and Vision Computing, 105: 104047. https://doi.org/10.1016/j.imavis.2020.104047

[14] Singh, V., Kaushik, V.D. (2022). DTCWTASODCNN: DTCWT based weighted fusion model for multimodal medical image quality improvement with ASO technique & DCNN. Journal of Scientific & Industrial Research, 81(8): 850-858. https://doi.org/10.56042/jsir.v81i08.56203

[15] Srinivasan, E.M., Ramar, K., Suruliandi, A. (2011). Texture analysis using local texture patterns: A fuzzy logic approach. International Journal of Pattern Recognition and Artificial Intelligence, 25(5): 741-762. https://doi.org/10.1142/s021800141100883x

[16] Strand, J., Taxt, T. (1994). Local frequency features for texture classification. Pattern Recognition, 27(10): 1397-1406. https://doi.org/10.1016/0031-3203(94)90072-8

[17] Jeong, N., Choi, J., Lee, G., Park, J., Kim, K. (2022). Feature selection for SAR target discrimination and efficient two-stage detection method. Remote Sensing, 14(16): 4044. https://doi.org/10.3390/rs14164044

[18] Raitoharju, J., Kiranyaz, S., Gabbouj, M. (2016). Feature synthesis for image classification and retrieval via one-against-all perceptrons. Neural Computing and Applications, 29(4): 943-957. https://doi.org/10.1007/s00521-016-2504-4

[19] Xu, J., Zhang, X., Zhao, C., Geng, Z., Feng, Y., Miao, K., Li, Y. (2023). Improving fine-grained image classification with multimodal information. IEEE Transactions on Multimedia, 26(15209210): 2082–2095. https://doi.org/10.1109/tmm.2023.3291819

[20] You, D., Wang, Y., Tao, C., Chen, Z., Jin, S. (2026). Cross-modal image fusion via dual attention and Mamba. Expert Systems with Applications, 310: 1131303. https://doi.org/10.1016/j.eswa.2026.131303

[21] Ansari, M.D., Ghrera, S.P., Mishra, A.R. (2016). Texture feature extraction using intuitionistic fuzzy local binary pattern. Journal of Intelligent Systems, 29(1): 19-34. https://doi.org/10.1515/jisys-2016-0155

[22] Shamir, L., Macura, T., Goldberg, I.G. (2010). Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. ACM Transactions on Applied Perception, 7(2): 1-17. https://doi.org/10.1145/1670671.1670672

[23] Wang, H., He, Z., Huang, Y., Chen, D., Zhou, Z. (2017). Bodhisattva head images modeling style recognition of Dazu Rock Carvings based on deep convolutional network. Journal of Cultural Heritage, 27: 60-71. https://doi.org/10.1016/j.culher.2017.03.006