Art Painting Style Transfer Algorithm Based on Deep Learning and Visual Attention Mechanism

Chuting Dai[1], Xiang Zhou[2]
[1]College of Fine Art Education, Guangxi Arts University, Nanning 530007, Guangxi, China
[2]College of Plastic Arts, Guangxi Arts University, Nanning 530007, Guangxi, China
Corresponding author: Xiang Zhou (e-mail: 15677193161@163.com)

ABSTRACT Robust semantic feature preservation and adaptive multi-scale information fusion are critical challenges in intelligent visual information processing and communication-enabled perception systems. To address structural distortion and semantic inconsistency in complex image style transfer, this study proposes a dual visual attention framework that integrates spatial attention and channel attention for adaptive feature enhancement and style-aware representation learning. A pre-trained VGG-19 network is employed to extract hierarchical semantic features, while the spatial attention module strengthens foreground structural preservation through semantic guidance and the channel attention module dynamically selects discriminative style representations using global feature responses. The weighted features are collaboratively fused through adaptive instance normalization and reconstructed by an end-to-end decoder under the joint optimization of content loss, style loss, and region consistency loss. Experimental results demonstrate that the proposed method achieves superior structural similarity, semantic preservation, and style fidelity compared with representative baseline algorithms, maintaining an SSIM above 0.85 under complex scenarios while significantly reducing style distortion and improving image quality. Beyond artistic image synthesis, the proposed architecture provides an effective framework for semantic information fusion, adaptive feature transmission, intelligent visual perception, and multi-scale representation learning, offering valuable methodological references for communication-oriented sensing systems, image-guided information processing, and engineering applications related to Electromagnetic Waves, Antennas and Propagation.

INDEX TERMS semantic information fusion, adaptive feature transmission, intelligent visual perception, multi-scale representation learning

I. INTRODUCTION

BREAKTHROUGHS in deep learning technology, have enabled style transfer to evolve from traditional non-realistic rendering to a data-driven end-to-end generation paradigm, showing broad application prospects [1,2]. However, how to ensure that algorithms accurately preserve the main semantic structure of the original image while mimicking the artist's brushstrokes has become a key challenge restricting further improvement in generation quality.

Major obstacles faced by contemporary investigations into style transfer are the maintenance of content and style balance. Formally, an ideal balance is defined as the state where the stylized image I_{cs} minimizes the total loss L_{total} such that the Structural Similarity Index (SSIM) between I_{cs} and the content image I_c remains above a threshold $\tau \approx 0.85$, while the style loss L_{style} is reduced to a local minimum that captures the characteristic Gram matrix statistics of the style image I_s without inducing semantic erosion. The vast majority of methods implemented for image style transfer focus predominantly on aligning the global average feature statistics of a particular image. This is done by treating each image as if it were an entirely homogeneous area for texturizing a particular image, thereby being incapable of semantically differentiating between the foreground subject and background area [3,4]. The results of applying expressive art style transfers will frequently see dense strokes in the background mistakenly rendered onto faces and buildings, resulting in the outline of incorrectly rendered content, thus losing some of the semantic information associated with content [5,6]. Artistic styles consist of many multiple-layered attributes, including but not limited to unique semantics (such as color tone, direction of brush strokes, texture features, etc.). Current methods do not have an effective means of decoupling the stylistic attributes of artistic work of art at varied levels of granularity, and they do not include the implementation of a differentiated method of style transfer [7,8]. The great majority of the efforts to introduce attention mechanisms to enhance the quality

of transfer of style from a model to a finished piece of art occur only at the level of weighting shallow features; a structurally dedicated module for including attention to enhance the semantic perception needs associated with the evaluation of fine art are generally not implemented, which results in lower interpretability and transfer accuracy to the models [9,10]. There is an urgent need to introduce transfer algorithms based on deep learning and visual attention mechanisms to further improve the quality of style transfer synthesized works.

To address the aforementioned challenges, the academic community has conducted extensive research. Xu proposed a neural style transfer algorithm based on pre-trained VGG network and Gram matrix matching, pioneering a research paradigm for deep style transfer, but its iterative optimization method has low computational efficiency [11]. Satchidanandam et al. proposed a perceptual loss network, which provides an improvement strategy by combining semantic segmentation and perceptual functional loss [12]. Li et al. proposed an adaptive instance normalization layer, which directly aligns the mean and variance of style features, realizing single-network style transfer [13]. Bai et al. proposed GCSANet (Global Context Self-Attentional Network) based on style attention network, which uses self-attention mechanism to perform local matching of style features, improving the precision of stylization [14]. Ding and Shi proposed a linear style transfer network, which achieves multi-style fusion by decoupling content structure and style pattern [15]. An et al. proposed the ArtFlow model, which introduces reversible normalization flow to preserve the integrity of content information [16]. Fengxue et al. constructed a style transfer framework based on Transformer, which uses long-distance dependency modeling capability to improve style consistency [17]. The aforementioned studies have advanced the field from the perspectives of computational efficiency, feature alignment, and model structure, but they have not fundamentally solved the problem of foreground structure distortion caused by a lack of semantic understanding. Semantic misalignment and content erosion phenomena still exist in the transfer of complex art styles. Therefore, how to design a visual attention mechanism that can spontaneously learn semantic regions and implement differentiated stylization has become a key problem that urgently needs to be solved in this field.

To resolve the structural distortion inherent in deep style transfer, this paper proposes a dual visual attention network. A pre-trained VGG-19 network is used to extract features from both the content and style images. A spatial visual attention module is constructed, which generates pixel-level response weight maps of the content image through a self-learning mechanism, guiding the network to spontaneously focus on the foreground region and impose stronger content constraints during the stylization process. Simultaneously, a channel-based style visual attention module is designed, applying global average pooling to the style feature map and calculating the activation response coefficients of different feature channels through a fully connected layer, achieving adaptive selection and recombination of global color features and local texture features. The weighted feature maps output by the two attention modules are collaboratively fused through an adaptive instance normalization layer, and finally reconstructed into a stylized image by a decoding network. During end-to-end training, the entire network jointly optimizes content loss, style loss, and region consistency loss. This loss is calculated by determining the difference between the generated image and the content image in semantic segmentation results, providing effective supervision signals for the spatial visual attention module. This paper innovatively integrates spatial visual attention and channel visual attention in parallel to achieve differentiated transfer of foreground protection and style texture; designs a region consistency loss function to provide supervision and constraints for the visual attention module from a semantic level; and constructs an attention feature fusion strategy to ensure that the foreground region is always subject to strong content constraints during the stylization process, thus systematically solving the semantic distortion problem in art style transfer from an algorithmic level.

II. STYLE TRANSFER NETWORK BASED ON DUAL VISUAL ATTENTION MECHANISM

A. OVERALL NETWORK ARCHITECTURE DESIGN

The style transfer network proposed in this paper, based on a dual visual attention mechanism, adopts an end-to-end encoder-decoder structure and consists of four functionally integrated parts: a feature extraction module, a dual attention module, a feature co-fusion module, and an image reconstruction module. As illustrated in Figure 1, the co-fusion module acts as the critical interface that aligns spatial and channel features via AdaIN before passing them to the decoder-based reconstruction module. The overall architecture of this transfer network is shown in Figure 1.

Figure 1 shows a feature extraction module based on a VGG-19 network pre-trained on the ImageNet dataset. It performs forward propagation on the input content image and style image respectively to obtain multi-level feature representations. The content image passes through higher layers of the network to obtain a feature map containing semantic information, while the style image passes through different layers to obtain a set of feature maps containing texture and color information. The dual attention module consists of a spatial visual attention branch and a channel style visual attention branch running in parallel. The spatial attention branch receives the content feature map and, combined with semantic segmentation priors, generates attention weights for the foreground region, enhancing the feature response of that region. The channel style attention branch receives the style feature map and, by calculating the activation coefficients of each feature channel, adaptively selects the texture features most relevant to the target style. The output features of the two branches are aligned and merged in the feature co-fusion module through an adaptive

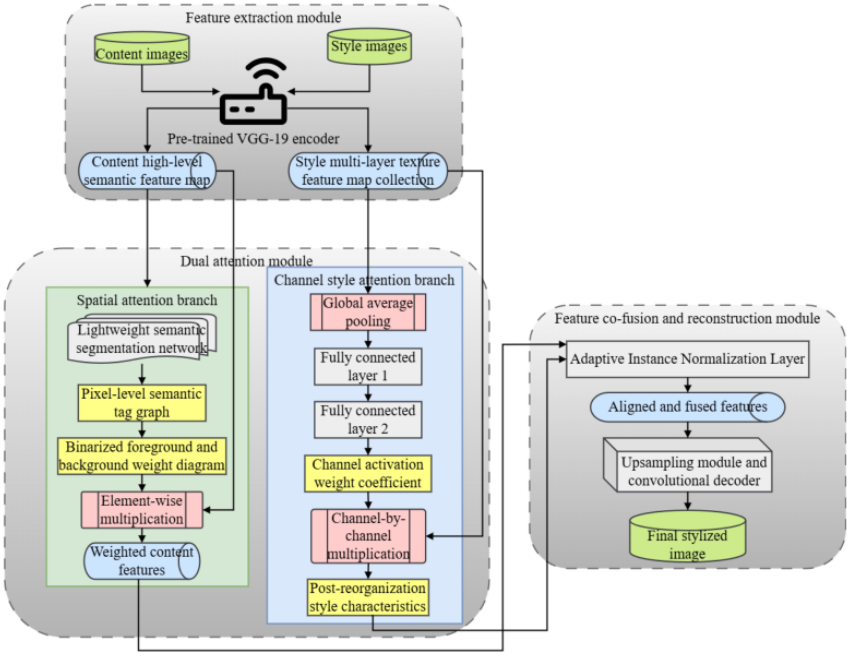

FIGURE 1. Network architecture for the transfer of artistic painting styles based on dual visual attention

instance normalization layer. Finally, the input is given to a decoder consisting of multiple upsampling modules and convolutional layers, progressively reconstructing a stylized image at a specified resolution. The entire network does not rely on any post-processing operations during training, achieving end-to-end mapping from content and style images to the generated image.

B. MULTI-LEVEL FEATURE EXTRACTION

This paper uses a VGG-19 network pre-trained on the ImageNet dataset as the feature extraction backbone, and performs forward propagation on the input content image and style image respectively to obtain multilevel feature representations for subsequent attention mechanisms and loss calculations [18]. The VGG-19 network contains 16 convolutional layers and 5 max pooling layers, and its hierarchical structure naturally corresponds to the progressive representation capability of images from pixel-level details to semantic-level abstraction.

Let the input content image be I_c and the input style image be I_s. Input I_c and I_s into the VGG-19 network respectively, and select the output of a specific network layer as the feature map. For content feature extraction, the output of the high-level convolutional layer conv4_2 of the network is selected as the content semantic feature map F_c. This layer of feature map has a large receptive field and can encode the structural information and spatial layout of objects in the image, ignoring pixel-level detail changes, so it is suitable for constraining the content structure of the generated image. The dimension of F_c is $C_c \times H_c \times W_c$, where C_c is the number of channels, and H_c and W_c are the height and width of the feature map, respectively. For style feature extraction, the outputs of convolutional layers from multiple network levels are selected to form a set of style feature maps. Specifically, the outputs of five layers—conv1_2, conv2_2, conv3_2, conv4_2, and conv5_2—are selected as the style feature representations $\{F_s^1, F_s^2, F_s^3, F_s^4, F_s^5\}$. Different layers correspond to style information at different scales. Shallow feature maps encode local texture, brushstroke details, and color distribution of the image, while deep feature maps encode more macroscopic compositional patterns and style tones. Each style feature map F_s^l has dimensions of $C_s^l \times H_s^l \times W_s^l$, where the superscript l indicates the network layer index.

In the subsequent style loss calculation, the Gram matrix is used to measure the relevance of style features

In the subsequent style loss calculation, the Gram matrix is used to measure the relevance of style features [19,20]. For any style feature map F_s^l, its spatial dimensions are flattened into a two-dimensional matrix $\tilde{F}_s^l$ with dimensions of $C_s^l \times (H_s^l W_s^l)$. The formula for calculating the Gram matrix G_s^l is:

$$G_s^l = \tilde{F}_s^l (\tilde{F}_s^l)^T \tag{1}$$

Matrix G_s^l has the dimension of $C_s^l \times C_s^l$ and each feature channel's pairwise correlation is contained in one element of this Gram matrix; thus, this cross-correlation statistic of all feature maps effectively describes the distribution of the texture's style, regardless of the spatial position of the feature maps. Therefore, it is commonly used to model styles.

While the Gram matrix discards explicit spatial coordinates to capture global texture, our framework reconciles this by feeding the content feature map F_c into the spatial

attention module to define region-specific boundaries. These spatial boundaries then act as a 'geometric container' for the global style statistics derived from the Gram matrices, ensuring that style textures are applied differentially—with higher content fidelity in the foreground and higher stylistic density in the background. Throughout the extraction process, all VGG-19 pre-trained parameters are frozen, meaning they will not receive any gradient updates, thus, maintaining the stability of feature representations and ensuring the consistency of the feature representations.

C. SPATIAL VISUAL ATTENTION MODULE

The spatial visual attention module's purpose is to provide various stylistic constraints to the different semantic areas of the content image in order to protect the structure of the foreground subject from having excessive erosion due to the background texture. The input to this module is the content feature map $F_c \in R^{C_c \times H_c \times W_c}$

output from the conv4_2 layer of a VGG-19 network. This produces a spatial attention weight map using a semantic segmentation prior and hence selectively enhancing the feature response of the foreground region. The semantic segmentation prior consists of a MobileNetV2-based DeepLabV3+ architecture, which is initially pre-trained on the MS-COCO dataset and subsequently fine-tuned on a synthetic 'stylized-natural' hybrid dataset to ensure robust feature extraction from both realistic and artistic textures, providing reliable supervision signals for the spatial visual attention module. The original content image I_c is passed through the segmentation network to produce a pixel level semantic label map $M \in R^{H \times W}$, in which each value $M_{ij} \in \{0, 1\}$ indicates if that pixel is part of the foreground subject (1) or background region (0) at pixel (i, j). The semantic label map is then downsampled to the same spatial size $H_c \times W_c$ as the content feature map F_c using bilinear interpolation. The downsampled semantic label map $M_{\text{down}} \in R^{H_c \times W_c}$ is obtained. The spatial attention weight graph $A_{\text{spatial}} \in R^{H_c \times W_c}$ is constructed based on the downsampled semantic label graph, and its calculation method is as follows:

$$A_{\text{spatial}}(i, j) = 1 + \alpha \cdot M_{\text{down}}(i, j) \tag{2}$$

Where α is the learnable foreground enhancement coefficient, initially set to 1.0. This unit initialization is chosen to ensure that, at the start of training, the foreground regions receive a 2x weighting boost $(1 + \alpha)$ compared to the background, providing a sufficiently strong gradient signal for the network to identify semantic importance. Empirical sensitivity tests indicated that while values between 0.5 and 2.0 all converge to similar final states, $\alpha = 1.0$ offers the most stable loss reduction curve in the first five epochs. Crucially, the spatial visual attention module also incorporates a series of learnable 1×1 convolutional layers following the semantic label map generation; these layers, alongside α, are optimized during training via backpropagation while the VGG-19 backbone remains frozen to maintain stable base features. This design ensures that the spatial location corresponding to the foreground region receives a weight greater than 1 $(1 + \alpha)$, while the background region retains a weight of 1. The weighted content feature map $\tilde{F}_c \in R^{C_c \times H_c \times W_c}$ is obtained through element-wise multiplication:

$$\tilde{F}_c(c, i, j) = F_c(c, i, j) \cdot A_{\text{spatial}}(i, j) \tag{3}$$

This operation amplifies the feature responses of the foreground region. During subsequent feature fusion and image reconstruction, the decoder receives stronger foreground feature signals, thus prioritizing the preservation of foreground structural details during stylization. Background region features remain unchanged, allowing for more complete transfer of style textures.

The spatial attention weight map, A_{spatial}, is dynamically generated in each forward propagation. Its computation depends entirely on the semantic content of the input image, achieving adaptive spatial attention allocation based on image content. The weighted content feature map, $\tilde{F}_c$, is the output of the spatial visual attention module and is passed to the outputs of the feature co-fusion module and the channel visual attention module for joint processing.

D. CHANNEL VISUAL ATTENTION MODULE

The goal of the channel visual attention module is to achieve adaptive selection and recombination of style features, enabling the network to dynamically adjust the contribution of different feature channels according to the attributes of the input style image. The module takes as input the set of style feature maps $\{F_s^1, F_s^2, F_s^3, F_s^4, F_s^5\}$ output from five selected layers (conv1_2, conv2_2, conv3_2, conv4_2, and conv5_2) of the VGG-19 network. By applying channel-dimensional attention weights to the features of each layer, it strengthens texture channels strongly correlated with the target style and suppresses interference from irrelevant channels.

For the style feature map of layer l, a global average pooling operation is first applied [21]. The feature map of each channel is compressed into a scalar value, obtaining the channel description vector $z^l \in R^{C_s^l}$. The formula for calculating global average pooling is:

$$z_k^l = \frac{1}{H_s^l \times W_s^l} \sum_{i=1}^{H_s^l} \sum_{j=1}^{W_s^l} F_s^l(k, i, j) \tag{4}$$

k is the channel index, and z_k^l represents the global response intensity of the k-th channel. This operation aggregates spatial dimensional information into channel-level statistics, preserving the overall activation level of each feature channel.

The channel description vector z^l is input to a channel attention subnetwork consisting of two fully connected

layers. This subnetwork learns to capture the non-linear dependencies between channels and outputs channel

activation weight coefficients $w^l \in R^{C_s^l}$. The computation process of the channel attention subnetwork is expressed as follows:

$$w^l = \sigma\left(W_2^l \cdot \delta(W_1^l \cdot z^l)\right) \tag{5}$$

W_1^l is the weight matrix of the first fully connected layer, compressing the number of channels to $1/r$ of the original dimension, where r is the decay ratio set to 16; δ represents the nonlinear transformation introduced by the ReLU activation function; W_2^l is the weight matrix of the second fully connected layer, restoring the number of channels to the original dimension; σ represents the Sigmoid activation function, mapping the output value to the interval between 0 and 1, forming the attention weight coefficients for each channel. This bottleneck structure design reduces the number of parameters while enhancing the ability to model interchannel dependencies.

The obtained channel weight coefficients w^l are multiplied channel-by-channel with the original style feature map F_s^l to obtain the weighted style feature map $\tilde{F}_s^l \in R^{C_s^l \times H_s^l \times W_s^l}$.

$$\tilde{F}_s^l(k,i,j) = F_s^l(k,i,j) \cdot w_k^l \tag{6}$$

This operation allows feature channels with high weighting coefficients to contribute more in subsequent processing, while channels with low weighting coefficients are effectively suppressed. Different image styles activate different channel combinations; for example, styles focused on color representation strengthen color-related channels, while styles focused on brushstroke texture strengthen edge-direction-related channels. The weighted style feature map $\tilde{F}_s^l$ from each layer is used as the output of the channel visual attention module and is passed to the outputs of the feature co-fusion module and the spatial visual attention module for joint processing.

E. FEATURE CO-FUSION AND IMAGE RECONSTRUCTION

The feature co-fusion module receives the weighted content feature map $\tilde{F}_c$ output by the spatial visual attention module and the weighted style feature maps $\{\tilde{F}_s^l\}$ of each layer output by the channel style attention module. It achieves co-alignment and fusion of the two features through an adaptive instance normalization layer. The adaptive instance normalization layer aligns the channel-level mean and variance of the content features with the corresponding statistics of the style features [22,23]. Its calculation process is as follows:

$$\mathrm{AdaIN}(\tilde{F}_c, \tilde{F}_s^l) = \sigma(\tilde{F}_s^l)\frac{\tilde{F}_c - \mu(\tilde{F}_c)}{\sigma(\tilde{F}_c)} + \mu(\tilde{F}_s^l) \tag{7}$$

$\mu()$ and $\sigma()$ calculate the mean and standard deviation of the feature map in the spatial dimension, respectively. This operation allows the fused feature map to retain the spatial structure of the content features while inheriting the global color statistical properties of the style features.

For multiple weighted style feature maps $\{\tilde{F}_s^l\}$ from different network layers, adaptive instance normalization is performed with the weighted content feature map $\tilde{F}_c$ to obtain multiple fused feature maps $\{T^l\}$. The decoder consists of multiple upsampling modules and stacked convolutional layers. Each upsampling module uses nearest neighbor interpolation to double the spatial size of the feature map, followed by two 3×3 convolutional layers for feature refinement [24,25]. After four upsampling stages, the feature map is gradually restored to the original input image resolution $H \times W$, and finally, a 1×1 convolutional layer maps the number of channels to RGB three channels, generating the final stylized image I_{cs}.

Taking the work "Blooming Flowers" as an example, Van Gogh's watercolor painting is used as the style image, and the ink painting on paper is used as the content image. The style transfer result is shown in Figure 2.

F. LOSS FUNCTION DESIGN

The proposed method jointly optimizes three objective functions: content loss, style loss, and region consistency loss, and constructs the overall loss function through weighted summation. The content loss constraint calculates the Euclidean distance using the feature map output from the VGG-19 network's conv4_2 layer. The content loss is defined as:

$$\mathcal{L}_{\text{content}} = \frac{1}{C_c H_c W_c}\left\|\phi_{\text{conv4\_2}}(I_{cs}) - \phi_{\text{conv4\_2}}(I_c)\right\|_2^2 \tag{8}$$

$\phi_{\text{conv4\_2}}$ represents the feature extraction function of the conv4_2 layer of the VGG-19 network, I_{cs} is the generated image, and I_c is the content image. This loss minimizes the difference between the generated image and the content image in the semantic feature space.

The style loss constrains the generated image by selecting feature maps from five layers conv1_2 to conv5_2, calculating the Euclidean distance between the Gram matrices of each layer, and defining the style loss as:

$$\mathcal{L}_{\text{style}} = \sum_{l=1}^{5}\frac{1}{C_s^l C_s^l}\left\|G(\phi_l(I_{cs})) - G(\phi_l(I_s))\right\|_2^2 \tag{9}$$

$\phi_l()$ is the feature extraction function of the l-th layer, $G()$ is the Gram matrix calculation operation, and I_s is the style image. This loss makes the generated image approximate the target style in terms of texture relevance distribution.

Region consistency loss directly constrains the structural integrity of the foreground region of the generated image. The generated image and content image are respectively input into the same lightweight segmentation

network as the spatial visual attention module to obtain their respective semantic label maps M_c and M_{cs}. The

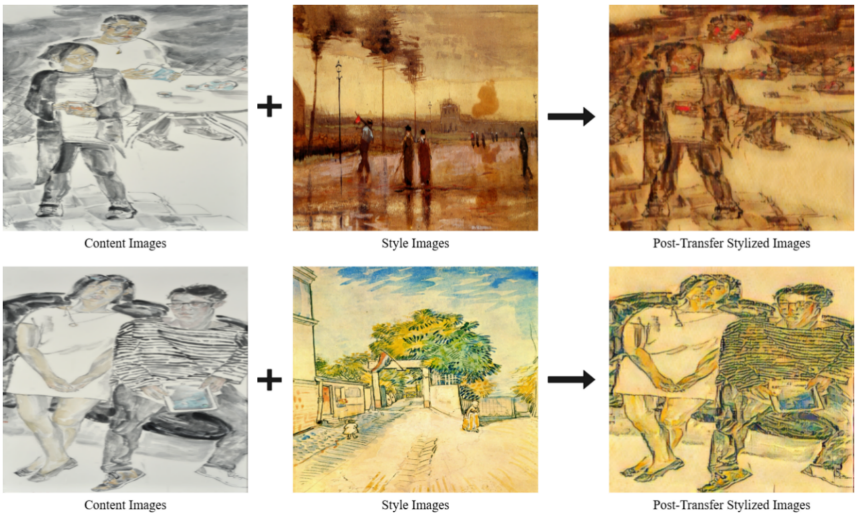

FIGURE 2. Example of painting style transfer

region consistency loss is defined as the cross-entropy between the two:

$$\mathcal{L}_{\text{region}} = -\frac{1}{N}\sum_{i,j}\left[M_c(i,j)\log M_{cs}(i,j) + (1-M_c(i,j))\log(1-M_{cs}(i,j))\right] \tag{10}$$

N is the total number of pixels, this loss function forces the foreground region of the generated image to maintain semantic consistency with the original image, thus protecting the main structure from stylistic erosion at the semantic level. The overall loss function $\mathcal{L}_{\text{total}}$ is the weighted sum of the three losses:

$$\mathcal{L}_{\text{total}} = \lambda_c\mathcal{L}_{\text{content}} + \lambda_s\mathcal{L}_{\text{style}} + \lambda_{\text{reg}}\mathcal{L}_{\text{region}} \tag{11}$$

λ_c, λ_s, and λ_{reg} are weighting coefficients balancing the various losses, set to 1.0, 10.0, and 5.0, respectively.

The above loss functions work synergistically with the dual attention module. Content loss and style loss guide the network to achieve basic style transfer, while region consistency loss provides semantic supervision to the spatial visual attention module, strengthening structural constraints on the foreground region. Together, they ensure that the generated image achieves accurate stylization while maintaining the semantic integrity of the content.

III. EXPERIMENTAL DESIGN

A. EXPERIMENTAL DATASET

This paper uses two public datasets as the content image library and the style image library, respectively. The content images are selected from the training set of the MS-COCO dataset. The natural images in this dataset cover a variety of semantic categories such as people, animals, buildings, and landscapes, which can fully verify the proposed algorithm's adaptability to different content images [26]. 5000 images covering primary categories (Person, Animal, Vehicle, Architecture, and Plant) are randomly selected from this dataset to form the training set. The 'difficulty levels' (1-5) mentioned in Section 4.1 are assigned based on the structural complexity and foreground-background contrast of the images, where level 1 represents high-contrast simple subjects and level 5 represents low-contrast cluttered scenes.

The style images are selected from the WikiArt dataset. This dataset contains paintings from hundreds of artists, covering a variety of art movements such as Impressionism, Expressionism, Abstract Art, and Realism. The image style features are significantly different, and the brushstrokes and textures are rich and diverse [27]. 2000 paintings are selected from this dataset to form the style image training set, and another 100 paintings are selected to form the style image test set. All images are uniformly adjusted to a resolution of 512×512 pixels before being input into the network, and the pixel values are normalized to the [0,1] interval. During

training, the content images and style images are randomly paired and input into the network to ensure that the model learns diverse style transfer mapping relationships.

B. SELECTION OF COMPARISON ALGORITHM

Four commonly used style transfer techniques were chosen as baseline models to test the effectiveness of the suggested technique(s).

1) AdaIN (Adaptive Instance Normalization) was chosen as the first comparison method because it applies adaptive instance normalisation to provide arbitrary stylisation content image by matching the statistics of both content and style sets at the channel level; making it the original real time style transfer architecture [28].

2) The SANet (Style Attention Network) method, based on style attention networks, introduces a self-attention mechanism to perform local matching of style features, improving the ability to model style texture details [29].

3) The Linear Style Transfer method, based on linear style transfer, achieves multi-style fusion by decoupling content

structure and style patterns, and can maintain the linear transformation characteristics of content structure [30].

4) The ArtFlow method, based on reversible normalized flow, uses a reversible network structure to preserve the integrity of content information, and has advantages in maintaining content structure.

All comparison algorithms were tested using default parameter settings on the same training set or using the pre-trained model provided by the official website to ensure the fairness of the comparison results.

C. EXPERIMENTAL PARAMETER SETTINGS

This paper uses the Adam optimizer to iteratively update the network parameters. Each iteration processes four pairs of content and style images simultaneously. The network training iterates through all training data once per epoch. The pre-trained parameters of the VGG-19 feature extraction network remain fixed during training and do not participate in gradient updates; only the parameters of the spatial visual attention module, channel visual attention module, and decoder network are updated. All experiments were conducted on a single NVIDIA Tesla V100 GPU using the PyTorch deep learning framework. Detailed experimental parameter configurations are shown in Table 1.

TABLE 1. Experimental Parameter Configuration Table

Parameter name	Value
Optimizer	Adam
Initial learning rate	0.0001
Batch size	4
Training epochs	20
Content loss weight λ_c	1.0
Style loss weight λ_s	10.0
Region consistency loss weight λ_{reg}	5.0
Input image resolution	512×512
VGG-19 network parameters	Fixed pre-trained weights
Deep learning framework	PyTorch
GPU model	NVIDIA Tesla V100

The above parameter configuration has been verified through preliminary experiments to achieve a good balance between generated image quality and training convergence speed. A learning rate of 0.0001 ensures the stability of the training process, and a batch size of 4 maximizes data diversity within the limitations of GPU memory. Twenty training epochs allow the loss function to decrease to a stable range.

IV. RESULTS AND ANALYSIS

A. CONTENT RETENTION ASSESSMENT

This section quantitatively assesses the spatial visual attention module's performance for maintaining accuracy to the overall design of the foreground through evaluating structural preservation ability across 5 levels of difficulty for content images. Structural similarity acts as an indicator to measure pixel-level structural similarity between the original content and generated images; therefore a higher score represents a greater degree of retention of original content. The average SSIM values of the four baseline algorithms and the proposed algorithm are calculated at each difficulty level on the test set. The comparison results of content preservation among different algorithms are shown in Figure 3.

In Figure 3, the horizontal axis represents different difficulty levels of the five content images, and the vertical axis represents the SSIM value. The data shows that the SSIM value of all algorithms decreases as the difficulty level of the content image increases. Our proposed algorithm achieves a higher SSIM value than the comparison methods at all difficulty levels, reaching 0.94 at difficulty level 1, a 2.2% improvement over ArtFlow's 0.92. At difficulty level 3, the proposed algorithm's SSIM value is 0.91, while AdaIN drops to 0.78, SANet to 0.82, Linear Style Transfer to 0.80, and ArtFlow to 0.84. At difficulty level 5, the proposed algorithm maintains an SSIM value of 0.85, ArtFlow drops to 0.71, and AdaIN is only 0.62. The reason the proposed algorithm maintains a high SSIM value at high difficulty levels is that the spatial visual attention module applies enhanced weights to the foreground region through semantic segmentation priors, causing the decoder to prioritize preserving foreground structural information during reconstruction and suppressing the erosion of the foreground by background texture. The previous comparison method lacked semantic awareness of image content, making it difficult to distinguish foreground from background in complex scenes, resulting in foreground structures being covered by style textures. Overall results show that the proposed algorithm has a significant advantage in content preservation.

B. STYLE SIMILARITY EVALUATION

To verify the adaptive selection capability of the channel visual attention module for style features, this section quantitatively evaluates the style transfer accuracy of different algorithms across five typical art styles. The style loss function value is used as an indicator to measure the consistency of texture statistical characteristics between the generated image and the target style image; a smaller value indicates a better style transfer effect. The average style loss values of the four baseline algorithms and the proposed algorithm are calculated for each style type on the test set. The comparison results of style similarity between different algorithms are shown in Figure 4.

Figure 4 shows the five art styles on the horizontal axis and the style loss value on the vertical axis. Error bars represent the standard deviation of five independent experiments. Our algorithm achieves the lowest style loss value across all style types, reaching 2.76 for Impressionism, a reduction of 0.52 compared to ArtFlow's 3.28 and 1.06 compared to AdaIN's 3.82. For Expressionism, the proposed algorithm achieves 2.98, SANet 3.78, and Linear Style Transfer 3.94, demonstrating a clear advantage. For Abstract art, the proposed algorithm's loss value is 3.21, ArtFlow 3.89, and AdaIN rises to 4.53. For Pointillism, the most

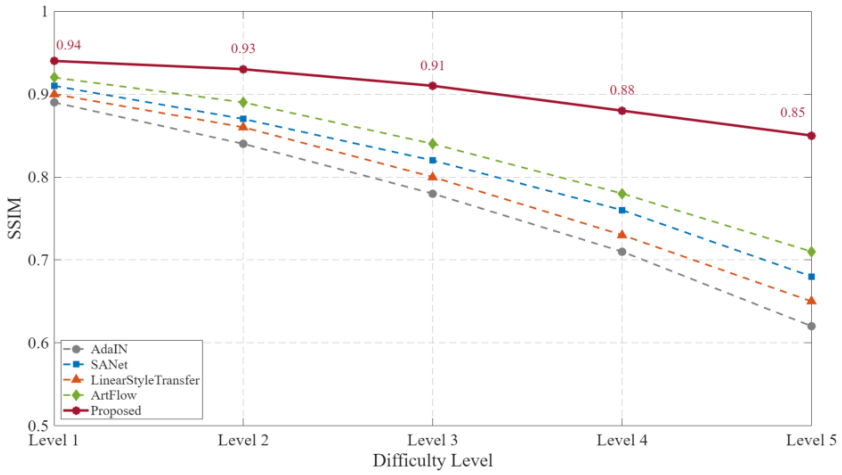

FIGURE 3. Comparison of content retention rates of different algorithms

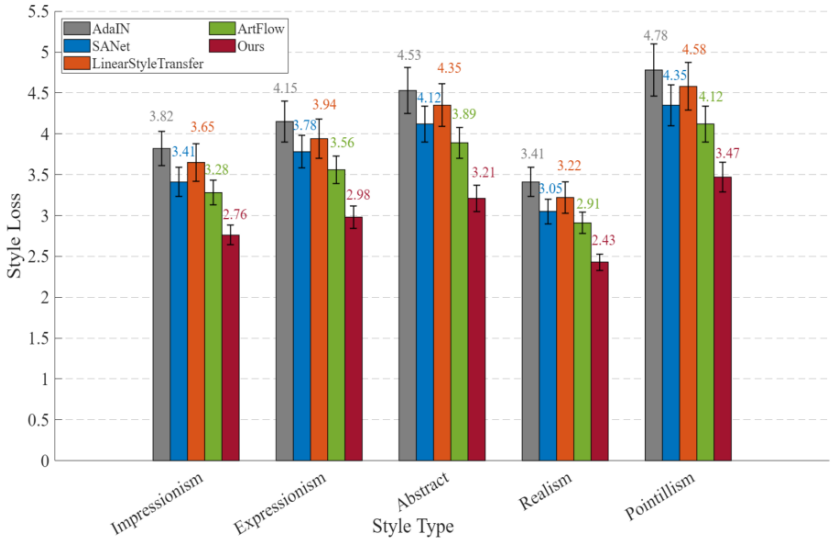

FIGURE 4. Comparison of loss styles of different algorithms

textured style, all algorithms show increased loss values, but the proposed algorithm maintains the lowest value of 3.47, a reduction of 0.65 compared to ArtFlow's 4.12 and 0.88 compared to SANet's 4.35.

The reason the proposed algorithm achieves a lower style loss value is that the channel-level style visual attention module adaptively selects between global color features and local brushstroke features by calculating the activation weight coefficients of each feature channel. This allows the network to dynamically enhance relevant texture channels and suppress irrelevant channels based on the attributes of the input style image. Contrastive methods do not utilize this feature-selection at the channel level, resulting in poorer accuracy when comparing style feature distributions after processing different types of styles, and hence increased amounts of style loss. The algorithm demonstrates lower style loss across various genres; furthermore, qualitative visual inspections (as seen in Figure 2 and subsequent user studies) confirm that the dual attention mechanism better captures the 'gestalt' of artistic brushstrokes compared to baseline methods, indicating that the quantitative reduction in loss aligns with a more authentic subjective artistic 'feel'.

C. COMPREHENSIVE EVALUATION OF GENERATION QUALITY

To verify the effect of the dual attention mechanism on improving the overall quality and realism of the generated images, this section comprehensively evaluates the FID (Fréchet Inception Distance) metric of different algorithms on the test set. The FID score calculates the distribution distance between the generated image set and the real image set in the Inception network feature space; a smaller FID score indicates higher quality and stronger visual realism of the generated image. Ten independent experiments were conducted for each algorithm, and the distribution characteristics of the FID scores were statistically analyzed. The comparison of the FID score distributions of different algorithms is shown in Figure 5.

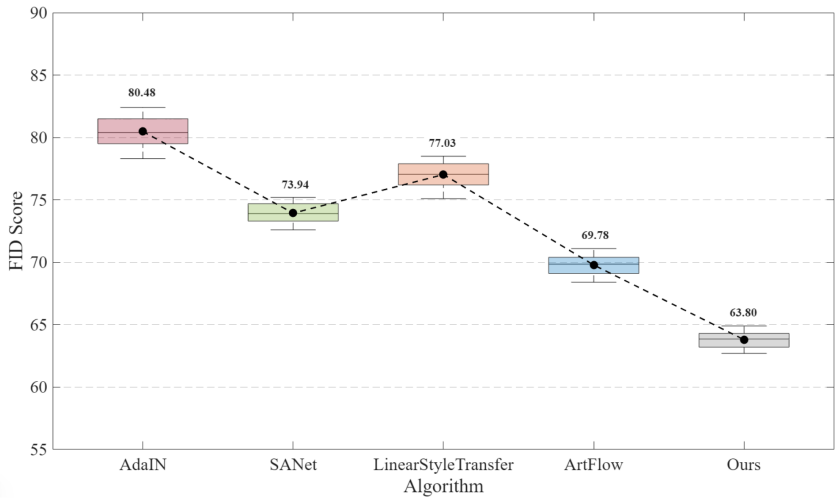

FIGURE 5. Comparison of FID score distributions for different algorithms

Figure 5 shows the score distribution range, median, quartiles, and mean of each algorithm in ten experiments. The average FID score of the algorithm presented in this paper is 63.80, significantly lower than the other four baseline algorithms. Furthermore, its score distribution is concentrated, ranging from 62.7 to 64.9, while AdaIN has the widest distribution range, reaching 78.3 to 82.4. The reason why the algorithm presented in this paper achieves the lowest FID score and the most concentrated distribution is due to the synergistic effect of the spatial visual attention module and the channel style attention module. The spatial visual attention component helps maintain the integrity of the foreground by preventing shifts in feature distributions resulting from the distortion of a given object, while the channel style attention component helps accurately select texture features that are representative of the target style so that the final generated image can be compared to the distribution of the original painting based on color and texture characteristics. Neither of the contrasting methods being compared has a dual attention mechanism. While AdaIN uses global alignment to create a distorted structure for the foreground, ArtFlow creates an image that maintains content characteristics but has a poor ability to match the texture of the target style, therefore resulting in a greater difference between the generated image and the original image based on how the feature distributions differ. Based on the results of experimental tests conducted using this approach, it can be concluded that this method provides substantial improvement in both image quality and stability.

D. SEMANTIC PRESERVATION ABILITY EVALUATION

The purpose of this Section is to verify the performance of various algorithms for semantic segmentation accuracy due to region consistency loss's impact on providing or maintaining the structural integrity of foreground regions. The average intersection-union ratio (IUGR) will be used as a measure of semantic consistency between the generated image with regards to the original content image. The generated image is input into a pre-trained semantic segmentation model, and the IUGR of the segmentation results is calculated between the generated image and the original content image. A higher IUGR indicates that the foreground region structure is better preserved. The IUGR of the five algorithms on the test set for the five main semantic categories is calculated, and the statistical results are shown in Table 2.

TABLE 2. Average Crossover Union Ratio of Each Algorithm Across Different Semantic Categories

Semantic categories	AdaIN	SANet	Linear Style Transfer	ArtFlow	Proposed
Figure	0.61	0.67	0.64	0.72	0.86
Animal	0.58	0.64	0.61	0.69	0.83
Architecture	0.64	0.70	0.67	0.74	0.88
Vehicle	0.59	0.65	0.62	0.70	0.84
Plant	0.56	0.62	0.59	0.67	0.81
Mean	0.596	0.656	0.626	0.704	0.844

As shown in Table 2, the average intersection-union ratio (IUR) of the proposed algorithm is higher than that of the comparative methods across all semantic categories. For the "person" category, it reaches 0.86, an improvement of 0.14 over ArtFlow and 0.19 over SANet. For the "building" category, the proposed algorithm reaches 0.88, while AdaIN only reaches 0.64. For the "animal," "vehicle," and "plant" categories, the proposed algorithm reaches 0.83, 0.84, and 0.81, respectively, while the suboptimal ArtFlow achieves 0.69, 0.70, and 0.67 for the same three categories. The reason for the higher semantic segmentation accuracy of the proposed algorithm lies in the fact that the region consistency loss directly constrains the consistency between the generated image and the content image at the semantic label level, while the cross-entropy loss forces the pixel-level classification results of the foreground region to remain consistent with the original image. The spatial visual attention module enhances the foreground response at the feature level, and the region consistency loss imposes semantic constraints at the loss function level. Together, these two mechanisms enable the decoder to prioritize the preservation

of the semantic structure information of the foreground during reconstruction. The contrasting methods lack this kind of supervisory signal for the attention mechanism. The global statistical alignment methods of AdaIN and Linear Style Transfer result in the foreground structure being covered by style textures. Although ArtFlow preserves content information through reversible flow, it does not apply differential protection to semantic regions. Therefore, the semantic segmentation accuracy of these methods is lower than that of the proposed algorithm proposed in this paper. The results show that the algorithm proposed in this paper has a significant advantage in semantic preservation.

V. CONCLUSION

To resolve the foreground structural distortion in deep style transfer caused by insufficient semantic understanding, this paper proposes a dual visual attention network optimized for preserving the intricate fiber morphology and yarn spatial distribution within industrial textures. This mechanism ensures that the synthesis of artistic patterns maintains the precise topographical characteristics of the fabric substrate while accurately mapping complex stylistic elements. The algorithm applies enhanced weights to the foreground region of the content image through a spatial visual attention module, adaptively selects style texture features through a channel visual attention module, and achieves co-optimization of foreground structure and style texture under the supervision of region consistency loss. Experimental results show that the proposed method achieves an SSIM value higher than 0.85 in content preservation, a loss value of 2.76 for Impressionist style similarity, and an average intersection-union ratio of 0.844 in semantic preservation, outperforming the comparison methods. This method effectively solves the semantic distortion problem in art painting style transfer, and can achieve accurate transfer of complex art textures without destroying the main structure of the content, providing technical support for digital art creation, digital preservation of cultural heritage, and intelligent image editing.

AUTHOR CONTRIBUTIONS

Conceptualization –Chuting Dai and Xiang Zhou; methodology – Chuting Dai and Xiang Zhou; investigation – Chuting Dai and Xiang Zhou; writing-original draft preparation – Chuting Dai and Xiang Zhou. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] S. Zhang, Y. Qi, and J. Wu, "Applying deep learning for style transfer in digital art: enhancing creative expression through neural networks," Scientific reports, vol. 15, no. 1, pp. 11744-11749, 2025, doi: 10.1038/s41598-025-95819-9.

[2] R. Bian, "Landscape Painting Style Transfer and Feature Extraction Model based on Convolutional Neural Networks," International Journal of Multiphysics, vol. 18, no. 2, pp. 581-586, 2024.

[3] Y. Yu, J. Wang, and N. Li, "Foreground and background separated image style transfer with a single text condition," Image and Vision Computing, vol. 143, no. 1, pp. 104956-104962, 2024, doi: 10.1016/j.imavis.2024.104956.

[4] N. Qi, Y. Li, R. Fu, et al., "STRAT: Image style transfer with region-aware transformer," Neurocomputing, vol. 617, no. 1, pp. 129039-129044, 2025, doi: 10.1016/j.neucom.2024.129039.

[5] W. Ye, X. Zhu, Z. Xu, et al., "A comprehensive framework of multiple semantics preservation in neural style transfer," Journal of Visual Communication and Image Representation, vol. 82, no. 1, pp. 103378-103382, 2022, doi: 10.1016/ j.jvcir.2021.103378.

[6] Y. Qin, "Enhancing media image style transfer with advanced StyleGAN2 architectures," Scientific Reports, vol. 16, no. 1, pp. 583-589, 2025, doi: 10.1038/s41598-025-30170-7.

[7] Y. Yu, J. Xing, and N. Li, "Multi-layer feature fusion based image style transfer with arbitrary text condition," Signal Processing: Image Communication, vol. 132, no. 1, pp. 117243-117251, 2025, doi: 10.1016/j.image.2024.117243.

[8] C. Hu, S. Sun, P. Zhao, et al., "Dual-domain style transfer network," Science Progress, vol. 108, no. 2, Art. no. 368504251325649, 2025, doi: 10.1177/00368504251325649.

[9] Y. Dong, S. Fan, L. Zheng, et al., "AttenStyler: Text-image style transfer based on attention mechanism," Neurocomputing, vol. 638, no. 1, pp. 130075-130082, 2025, doi: 10.1016/j.neucom.2025.130075.

[10] H. Peng, W. Qian, J. Cao, et al., "Arbitrary style transfer based on attention and covariance-matching," Computers & Graphics, vol. 116, no. 1, pp. 298-307, 2023, doi: 10.1016/j.cag.2023.08.029.

[11] Y. Xu, "CNN-based image style transformation–Using VGG19," Appl. Comput. Eng, vol. 39, no. 1, pp. 130-136, 2024, doi: 10.54254/2755-2721/39/20230589.

[12] A. Satchidanandam, S. Al Ansari R M, L. Sreenivasulu A, et al., "Enhancing style transfer with GANs: perceptual loss and semantic segmentation," Int. J. Adv. Comput. Sci. Appl, vol. 11, no. 1, pp. 321-329, 2023, doi: 10.14569/IJACSA.2023.0141132.

[13] H. Li, L. Wang, and J. Liu, "Application of multi-level adaptive neural network based on optimization algorithm in image style transfer," Multimedia Tools and Applications, vol. 83, no. 29, pp. 73127-73149, 2024, doi: 10.1007/s11042-024-18451-1.

[14] Z. Bai, H. Xu, X. Zhang, et al., "GCSANet: Arbitrary style transfer with global context self-attentional network," IEEE Transactions on Multimedia, vol. 26, no. 1, pp. 1407-1420, 2023, doi: 10.1109/TMM.2023.3282489.

[15] S. Ding and Y. Shi, "Optimization Method for Controllable Image Style Transfer Based on DDPM Guided by Cross-Modal Text," IEEE Access, vol. 13, no. 1, pp. 210246-210256, 2025, doi: 10.1109/ACCESS.2025.3643194.

[16] J. An, S. Huang, Y. Song, et al., "ArtFlow: Unbiased Image Style Transfer via Reversible Neural Flows," Conference on Computer Vision and Pattern Recognition, vol. 21, no. 03, pp. 862-871, 2021, doi: 10.1109/CVPR46437.2021.00092.

[17] S. Fengxue, S. Yanguo, L. Zhenping, et al., "Image and video style transfer based on transformer," IEEE Access, vol. 11, no. 1, pp. 56400-56407, 2023, doi: 10.1109/ACCESS.2023.3283260.

[18] X. Cheng, F. Wang, A. Siddique A, et al., "Implementation of Neural Style Transformation Technique for Artistic Image Processing Using VGG19," IET Software, vol. 2025, no. 1, pp. 4145192-4145197, 2025, doi: 10.1049/sfw2/4145192.

[19] D. Jin, Z. Jin, Z. Hu, et al., "Deep learning for text style transfer: A survey," Computational Linguistics, vol. 48, no. 1, pp. 155-205, 2022, doi: 10.1162/coli_a_00426.

[20] H. Chen, F. Shao, X. Chai, et al., "Quality evaluation of arbitrary style transfer: Subjective study and objective metric," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 7, pp. 3055-3070, 2022, doi: 10.1109/ TCSVT.2022.3231041.

[21] H. Chen, F. Shao, X. Chai, et al., "Collaborative learning and style-adaptive pooling network for perceptual evaluation of arbitrary style transfer," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 11, pp. 15387-15401, 2023, doi: 10.1109/TNNLS.2023.3286542.

[22] X. Jin, C. Lan, W. Zeng, et al., "Style normalization and restitution for domain generalization and adaptation," IEEE Transactions on Multimedia, vol. 24, no. 1, pp. 3636-3651, 2021, doi: 10.1109/TMM.2021.3104379.

[23] Q. Zhang, S. Wang, and D. Cui, "Feature consistency-based style transfer for landscape images using dual-channel attention," IEEE Access, vol. 12, no. 1, pp. 164018-164027, 2024, doi: 10.1109/ACCESS.2024.3485063.

[24] I. Hwang and T. Oh, "Design and experimental research of on device style transfer models for mobile environments," Scientific Reports, vol. 15, no. 1, pp. 13724-13729, 2025, doi: 10.1038/s41598-025-98545-4.

[25] Huang C , Zhang J , Yuan H .Arbitrary Clothing Style Transfer Based onAttention Mechanism.Lecture Notes in Computer Science, 2025:239-252, doi: 10.1007/978-3-031-78125-4_17.

[26] Dong Y , Fan S , Zheng L ,et al.AttenStyler: Text-image style transfer based on attention mechanism.Neurocomputing, 2025(Jul.14):638, doi: 10.1016/j.neucom.2025.130075.

[27] W. Wang, H. Li, R. Sulaiman, et al., "Research on artistic style recognition and image transfer method based on deep visual feature extraction," International Journal of Information and Communication Technology, vol. 26, no. 40, pp. 86-103, 2025, doi: 10.1504/IJICT.2025.149816.

[28] Y. Zhang, B. Hu, Y. Huang, et al., "Adaptive style modulation for artistic style transfer," Neural Processing Letters, vol. 55, no. 5, pp. 6213-6230, 2023, doi: 10.1007/s11063-022-11135-7.

[29] Yuan X .Research on Painting Style Transfer Algorithm Based on Non-Negative Matrix Factorization.Advances in Transdisciplinary Engineering, 2024, doi: 10.3233/atde240507.

[30] J. Gao, Y. Hua, G. Hu, et al., "Discrepancy-guided domain-adaptive data augmentation," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 8, pp. 5064-5075, 2021, doi: 10.1109/TNNLS.2021.3128401.