IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 October 2023, accepted 3 December 2023, date of publication 13 December 2023,
date of current version 26 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3342444

RESEARCH ARTICLE

AI-Enhanced Digital Creativity Design: Content-Style Alignment for Image Stylization

LANTING YU[1] AND QIANG ZHENG[2]
[1]School of Publishing and Media, Chongqing Business Vocational College, Chongqing 401331, China
[2]CISDI Info, Chongqing 401122, China

Corresponding author: Lanting Yu (racenbeymoroo@gmx.com)

ABSTRACT This paper presents an AI(Artificial Intelligence)-powered method for enhancing digital creative design through image stylization. To achieve this, we introduce the Content-Style Alignment Module (CSAM), which includes the Dual-Stream Content-Style Processing Block (DS-CSPB), Content-Style Matching Attention Block (CS-MAB), and Content-Style Space-Aware Interpolation Block (CS-SAIB). DS-CSPB removes style information from content descriptors using whitening transformation while preserving semantic structures. CS-MAB reorganizes each content descriptor with its most relevant style descriptor, ensuring optimal style adaptation for content semantics. CS-SAIB aligns content and style descriptors in the same space, enabling diverse semantic distributions in content images to match various style patterns. Moreover, we introduce the Multifaceted Optimization Loss (MOL). This loss comprises multiple components: The relaxed Earth Mover Distance (rEMD) loss enhances color and texture distributions on content images. The Moment Matching (MM) loss reduces visual artifacts caused by cosine distance. The differentiable Color Histogram (CH) loss efficiently addresses color blending issues, preserving image naturalness. The content loss ensures no significant deformation or distortion during stylization. The reconstruction loss constrains all encoder-decoder features to the VGG feature space, maintaining shared spaces between content and style descriptors. We conducted extensive comparative and ablation experiments, which demonstrated superior performance in image stylization, resulting in high-quality stylized images. Additionally, we provide a comprehensive review of current research in image stylization, effectively bridging the gap in this area.

INDEX TERMS Deep learning, stylization, encoder-decoder structure, VGG.

I. INTRODUCTION

Image stylization, a technique in the fields of computer graphics and computer vision, aims to transform one image's appearance and artistic style into another [1], giving the image a unique artistic sense, as depicted in Figure 1. This process commonly employs machine learning algorithms to blend one image's content with another's style, resulting in a new image that retains the original content while adopting a different artistic style. Image stylization has a broad range of applications in areas like art creation, image editing, and filmmaking. It enables the transformation of ordinary photos into various artistic styles, such as oil paintings, watercolors, impressionism, and more, thereby enhancing images with creativity and artistic value. Furthermore, image stylization can be employed in design, advertising, and the media industry to create captivating visual effects.

Traditional image stylization methods are often considered within the broader context of texture synthesis, where textures are extracted from a style image and transferred to a content image. Efros et al. [2] employed a Markov Random Field (MRF) model to select the nearest neighborhood texture segment for filling in pixel values, a classic but computationally expensive approach requiring iterations over texture segments for each pixel. Wei et al. [3] improved the speed of texture synthesis through vector quantization. Ashikhmin et al. [4] introduced texture synthesis algorithms suitable for natural landscapes. Liang et al. [5] proposed algorithms that created mixed texture images by referencing multiple texture sample images. Han et al. [6] introduced a

The associate editor coordinating the review of this manuscript and approving it for publication was Taous Meriem Laleg-Kirati.

new sample-based multi-scale texture synthesis algorithm, enabling texture synthesis for low-resolution images at different scales.

Despite the usefulness of traditional stylization techniques in generating various artistic works, they still have limitations: (1) Limited generalization to specific styles, (2) Focus on low-level details, neglecting high-level semantic features, and (3) Slowness in image transformations due to per-pixel computation. To overcome these limitations, neural network-based stylization techniques have emerged.

FIGURE 1. A rendering of neural style transfer by Gatys et al [1]. (a) Content image. (b) Style image. (c) Resulting image.

Gatys et al. [1] discovered that deep neural networks can simultaneously capture low-level texture information and high-level semantic information from images. They introduced this technology to the field of style transfer and pioneered a neural style transfer model based on the VGG (Visual Geometry Group) network, as shown in Figure 1. Gatys' work attracted widespread attention from both academia and the art world. However, their method, which simply blends style patterns from different feature layers onto the content image, overlooks the preservation of content semantics, resulting in stylized images that often suffer from blurry boundaries and distorted overall contours. To address this issue, a simple yet effective technique called Adaptive Instance Normalization (AdaIN) was proposed [7]. AdaIN achieves style transfer at the feature level by altering the distribution of features to support input from arbitrary style images. However, this method has limitations as it computes the mean and variance of features globally, ignoring local details, and therefore, it diminishes local stylization performance. While these methods can effectively transfer color and texture information from style images to target images, they lack local perception capabilities in stylization models, leading to the loss of local semantic information. This results in unnatural visual effects where detailed information in the target image is directly stylized and integrated with the overall semantic structure.

To improve the local perception capabilities of arbitrary stylization models, Park et al. [8] introduced Style Attentional Networks (SANet). SANet matches style descriptors with content descriptors and pays more attention to similar feature regions in style images. This method has been proven effective in generating more local stylistic details in arbitrary style transfer. However, it only re-weights feature maps of style images and simply integrates them into content descriptors during decoding. This causes the stylized style to blur around the boundaries of objects in the content image. Furthermore, it fails to align content descriptors with their associated style descriptors correctly, resulting in inappropriate color and texture information transfer from content semantics to artistic styles. Subsequent works of SANet, such as the methods proposed in [9] and [10], demonstrate the superior performance. These methods utilize learnable kernels and compute pairwise similarities to generate attention maps, serving as fine-grained, point-wise feature transformations for stylization. However, the aforementioned attention-based style transfer methods are not without limitations. Due to the inherent differences in content semantics and style semantics, their feature distributions are heterogeneous. As the result, the attention blocks struggle to learn the required feature matching techniques, leading to incoherent alignment between semantic regions and style descriptors. In other words, the same content semantic region may be presented in various different style patterns, resulting in visual artifacts and chaotic stylization outcomes.

Addressing the issues in existing methods, we propose a Content-Style Alignment Module (CSAM) to enhance image stylization. Specifically, to solve problems of overall contour distortion and the loss of local semantic information leading to the loss of detailed information, we apply whitening transformation and position-level feature processing to content descriptors. Additionally, to address the problem of misalignment between content descriptors and related style descriptors, an attention fusion block designed for artistic style transfer is introduced, allowing each content descriptor to be rearranged with its related style descriptor. Finally, to resolve the incoherence between content semantics and style descriptor matching, a space-aware interpolation block is introduced, enabling individual pairing of content image's semantic distribution and style image's stylistic pattern.

In summary, our main contributions are as follows:

(1) Introduction of CSAM, a network enabling one-to-one matching of content semantics and style patterns while preserving the semantic structure of content images.

(2) Proposal of a dual-stream feature processing module to maintain both the overall contour and local details of content images.

(3) Design of an attention feature fusion block addressing misalignment between content and style descriptors, ensuring appropriate transfer of content semantics to colors and textures.

(4) Presentation of a space-aware interpolation block to rectify incoherence between content semantics and style descriptor matching.

(5) Enhancement of the loss function, incorporating relaxed Earth Mover Distance (rEMD) loss for style feature optimization, Moment Matching (MM) loss to reduce visual artifacts, Color Histogram (CH) loss to control color blending, content loss for semantic preservation, and image reconstruction loss for feature space consistency.

(6) Extensive quantitative and qualitative experiments on benchmark style datasets and diverse content datasets, demonstrating the effectiveness, efficiency, and generality of our approach.

(7) A systematic and comprehensive review of the current landscape of image stylization, highlighting the strengths and weaknesses of representative methods, serving as a valuable reference for future researchers.

II. RELATED WORKS

This section provides a systematic overview of image stylization, including non-neural network-based image stylization and neural image stylization methods. It categorizes them in detail and systematically explains the algorithm principles and their advantages & disadvantages for representative models.

A. NON-NEURAL NETWORK IMAGE STYLIZATION

Artistic stylization has long been an important research area in computer graphics due to its widespread applications. Before the advent of deep learning-based image style transfer, the related research extended into the field of Non-Photorealistic Rendering (NPR). However, most NPR algorithms were designed for specific artistic styles, making it challenging to extend them to other styles. This section will briefly review some traditional image stylization algorithms.

1) STROKE-BASED RENDERING

Stroke-Based Rendering (SBR) is the process of rendering images with specific styles by placing discrete elements called strokes on a virtual canvas.

SBR algorithms aim to faithfully replicate a specified style and can effectively simulate certain types of styles, such as oil painting, watercolor, or sketching. However, most SBR algorithms are typically designed for specific styles and cannot replicate arbitrary styles.

2) ANALOGY-BASED IMAGE STYLIZATION

Using the analogy-based approach, Hertzmann et al. [11] synthesized images with new textures by mapping image feature relationships. Image analogy algorithms learn analogy transformations in sample training pairs and generate stylized images that are similar when given a test input photograph. Image analogy can also be extended in various ways, such as learning brush positions for portrait rendering.

In general, image analogy works well for various artistic styles but often lacks paired training data. Another limitation is that image analogy relies only on low-level features, making it ineffective at capturing image content and style, resulting in less-than-ideal synthesized image quality.

3) IMAGE FILTERING TECHNIQUES

The creation of artistic images aims to simplify and abstract the subject matter. Therefore, it is possible to use relevant image filters to render specific photos. Gooch et al. [12] first used the difference of bilateral filters and Gaussian filters to achieve cartoon-like effects.

Compared to other types of image stylization techniques, image filtering techniques are faster, more stable, and suitable for industrial applications. However, they are limited in terms of style diversity.

B. NEURAL STYLE TRANSFER METHODS

This section provides an overview of mainstream neural style transfer methods, including slow neural stylization based on image iteration and fast neural neural stylization based on model iteration. Slow stylization based on image iteration generates stylized images through pixel iteration on noisy images, further categorized into statistical and non-statistical parameter-based methods depending on the style matching approach. Methods based on non-parametric approaches primarily rely on region block similarity for style transfer, yielding better results when the content image closely resembles the style image in terms of shape. The second category of fast style transfer algorithms based on model iterations includes those using feed-forward models and methods based on Generative Adversarial Networks (GANs). Among these, algorithms based on feedforward stylization models achieve rapid style transfer by pretraining the generation model to stylize images. On the other hand, methods based on GAN networks primarily transform input image styles through the adversarial interplay between generators and discriminators. In summary, slow neural stylization methods based on image iterations achieve stylized images by iteratively processing pixels in the image, resulting in low computational efficiency. In contrast, fast neural stylization methods based on model iterations utilize generative models to stylize images, significantly improving processing speed. However, they also suffer from drawbacks, including poor generation quality and limited flexibility.

1) SLOW NEURAL STYLIZATION BASED ON IMAGE ITERATION

Slow neural stylization based on image iteration first extracts image features using deep neural networks and then iteratively updates the pixel values of noisy images using Convolutional Neural Networks (CNN). This process aligns the semantic features of the noisy image with the content descriptors of the content image and the style descriptors of the style image. The slow neural stylization based on image iteration defines two types of loss functions: content loss and style loss, with the style loss being the key component. It can be further divided into statistical and non-statistical parameter-based methods.

a: STATISTICAL PARAMETER-BASED METHODS

Statistical parameter-based methods use global statistical information for style transfer, such as Gram matrix-based methods and Maximum Mean Discrepancy (MMD)-based methods.

Gram matrix-based methods. The Gram matrix method was introduced by Gatys et al. in 2015 [13] and has been widely used to represent style descriptors. It minimizes the difference between the content and style descriptors of the generated image and the input image. However, Gatys et al.'s method fails to capture long-term correlations in images. Additionally, the use of Gram matrices to represent style descriptors has limitations in terms of stability and texture quality. Moreover, Gatys et al.'s method only extracts high-level image features, neglecting low-level information, which can result in the loss of fine details in stylized images. Furthermore, Gatys et al.'s algorithm does not consider factors like brushstroke variations, semantic information, and depth positioning in images, leading to unrealistic stylization. Subsequent algorithms have sought to address these shortcomings. Berger and Memisevic [14] improved upon Gatys' method by incorporating Markov structures into high-level features, enabling the generation of images that exhibit long-term consistency, suitable for generating textures with global symmetry and transforming image seasons. Risser et al. [15] discovered that the instability of Gram matrices primarily arises from their inability to capture the distribution information of image features. This can lead to different images with distinct data distributions having identical Gram matrices. To address this issue, Risser et al. [15] introduced an additional statistical histogram loss to represent the distribution information of image features, resolving the instability of Gram matrices. However, this algorithm is computationally complex. To address the problem of low-level information loss in content images, Li et al. [16] introduced a Laplacian loss to impose additional constraints on low-level features. They used a Laplacian matrix to describe low-level information in content images, complementing the high-level semantic information extracted from the VGG network. Subsequent research introduced semantic information to enhance control over generated images. Castillo et al. [17] incorporated instance-based semantic segmentation into Gatys et al.'s method to achieve style transfer for specific regions. Luan et al. [18] achieved style transfer in semantically matching subregions by manually controlling the mapping of semantic features between content and style images, preventing style overflow across different regions. Penhouest et al. [19] improved upon Luan et al.'s [18] approach by introducing automatic image semantic segmentation, simplifying the workflow.

MMD-based methods. Li et al. [20] proposed a new interpretation for neural style transfer, viewing it as a domain adaptation problem. They used MMD to compare the style differences between the source and target domains. By minimizing MMD, they reduced the domain gap, completing image stylization from the source domain to the target domain. Li et al.'s algorithm provided a mathematical explanation for the matching principle of Gram matrices, demonstrating that matching the Gram matrices of style images and generated images essentially minimizes the MMD between the two domain distributions. Therefore, various MMD algorithms with different kernel functions can be used for style transfer. This conclusion enhances the theoretical understanding of neural style transfer networks in academia.

Although the aforementioned methods have improved upon Gatys et al.'s algorithm, addressing issues such as instability, loss of details, and lack of semantic information, they have not yet resolved the problems related to brushstroke variations and the absence of depth positioning information. These issues remain significant factors affecting image generation quality.

b: NON-STATISTICAL PARAMETER-BASED METHODS

Non-statistical parameter-based methods, on the other hand, first segment both images into multiple regions and then match the most similar regions between the two images to achieve style transfer. These methods are effective in preserving local image features. They include MRF, semantic style transfer, and deep image analogy, based on deep neural networks. Each of these methods has its advantages and limitations.

Markov Random Fields. Li et al. [21] observed that early traditional style transfer methods based on MRFs only captured correlations between individual pixel features without constraining their spatial layout. Consequently, they proposed combining MRFs with dCNN (deep Convolutional Neural Network) for style transfer. They used a MRF model to segment the image feature maps extracted by dCNN, creating many regions, and matched regions between the two images by capturing feature information of local pixels.

However, Li et al.'s [21] algorithm does not yield satisfactory results when there is a significant difference between the content image and the style image. Additionally, the algorithm does not preserve image details and global semantic information effectively.

Semantic Style Transfer. Since Li et al.'s [21] algorithm does not perform precise mask segmentation of images, it can lead to incorrect semantic matches. Therefore, Champandard et al. [22] combined semantic segmentation with Li et al. 's MRFs algorithm to achieve semantic style transfer. However, the MRFs algorithm has high complexity. Hence, Chen et al. [23] introduced a new content-aware semantic mapping model to replace MRFs. This model uses masks to constrain the spatial correspondence between source and target images while incorporating high-order statistical information of style descriptors to enhance style matching consistency, simplifying the generation process. Subsequently, Merchrez et al. [24] proposed a new contextual loss, which considers only the similarity between image features, disregarding the spatial positions of features, enabling semantic style transfer without the need for spatial alignment.

Deep Image Analogy. Unlike MRFs-based methods, Liao et al.'s [25] deep image analogy finds the most similar regions between two images using the nearest-neighbor algorithm, aligning the features of the two images for style

transfer. However, this method does not effectively preserve the global semantic information in the images. Therefore, Gu et al. [26] introduced a feature rearrangement loss on top of this approach, adding constraints to the nearest-neighbor algorithm to match as many region blocks as possible, maintaining the global semantics of the image.

Although the aforementioned non-parametric methods have expanded the horizons of slow neural style transfer, they still have several limitations: (1) They require a certain degree of similarity in shape between the content and style images. (2) They lack effective representation of global semantic features in images. (3) They tend to produce images with relatively uniform style patterns, lacking richness in style descriptors.

2) FAST NEURAL STYLIZATION BASED ON MODEL ITERATION

While neural style transfer methods based on iterative image processing have achieved impressive results, their time-consuming nature, which requires iterations over each pixel in an image, has been a significant drawback. Johnson et al. [27] proposed a solution to address this issue by sacrificing flexibility in style selection and delegating the image generation process to a pre-trained feedforward stylization network, greatly enhancing the speed of style transfer. This section elaborates on model iteration-based methods, including those based on feedforward generation models and GAN-based approaches. In the case of feedforward generation models, a significant amount of image data is initially required to train the stylization network. Once trained, this network can directly produce stylized results from input content images. In contrast, GAN-based methods involve training both a generator and a discriminator until they reach a Nash equilibrium, at which point the generated images are indistinguishable from real ones.

a: FEEDFORWARD STYLIZATION MODEL-BASED METHODS

There are two representative works in the field of feedforward stylization model algorithms, namely, those by Johnson et al. [27] and Ulyanov et al. [28]. Both of these models share a common approach of stylizing images through pre-trained feedforward models, differing primarily in their model architectures. Johnson et al. [27] introduced residual blocks and strided convolutions on top of the model architecture proposed by Radford et al. [29], while Ulyanov et al. [28] utilized a multi-scale architecture for their generation network. Building upon Gatys et al.'s algorithm [13], Johnson et al. [27] pioneered fast style transfer using feedforward stylization models, achieving real-time style transformation. They introduced a perceptual loss function, which aligns with the two loss functions proposed by Gatys et al. [13]. Ulyanov et al. [28] introduced an image generation model with a multi-scale architecture, allowing it to learn features of the input image across different scales, resulting in more detailed generated images. Compared to Johnson et al. [20], Ulyanov et al.'s [28] model employed more parallel channels, reducing the model parameters and further enhancing style transfer speed. Subsequently, Ulyanov et al. [30] discovered that replacing batch normalization (BN), as used in their original model, with instance normalization (IN), which normalizes each image individually, significantly improved the quality of generated images. However, the above mentioned algorithms had limitations such as loss of low-level information, lack of stroke variation, and absence of deep positional information. Moreover, their generated images slightly lagged in quality compared to Gatys et al.'s [13] algorithm.

While the above generation model methods improved the speed of style transfer by two orders of magnitude compared to earlier iterative image stylization methods, they could only generate images in specific styles. Achieving other styles required retraining a feedforward generation network, which was inflexible and time-consuming. Thus, single-model multi-style generation networks emerged, integrating multiple styles into a single model to enhance the efficiency of feedforward networks. Dumoulin et al. [31] improved upon Ulyanov et al.'s [30] work by introducing Conditional Instance Normalization (CIN), enabling the creation of networks capable of transforming images into 32 different styles. Chen et al. [32] proposed the concept of StyleBank layers, where style descriptors are associated with a set of parameters in StyleBank layers, while content descriptors are shared. To achieve new style transfers, only a new StyleBank layer needs to be separately trained. These algorithms bound style descriptors to a small number of model parameters, reducing the training burden. However, as the number of learned styles increases, the model parameters become more redundant. Therefore, Zhang et al. [33] introduced a style selection model that incorporates multiple styles, using pixel values as input signals to control the generation of stylized images. This model can synthesize over 300 different textures and generate images in 16 different styles using a feedforward network. Zhang et al. [34] introduced the concept of CoMatch layers, where the model first learns multiple styles and then guides the input image to match the style descriptors in the CoMatch layer based on the target style image, achieving stylization.

Subsequently, Chen et al. [35] introduced the concept of arbitrary style transfer models. They used a pre-trained VGG network to extract multiple activation blocks from both content and style descriptors, and then matched each content activation block with the most similar style activation block, referred to as "Style Swap," to generate images. However, this model was slower in terms of style transfer speed. Huang et al. [36] improved upon Dumoulin's [31] work by enhancing CIN with the introduction of Adaptive Instance Normalization (AdaIN) layers, achieving real-time arbitrary style transformation for input images. However, AdaIN is data-driven and has limitations when generalizing to unseen data. Additionally, AdaIN only alters the mean and variance of feature maps, making it challenging to generate images with rich details and complex structures.

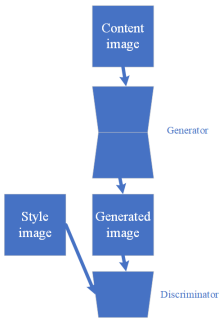

FIGURE 2. Sketch for image stylization using GAN.

Lu et al. [37] introduced semantic stylization into feedforward generation models, enabling arbitrary style transfer based on semantic correspondence. Li et al. [38] proposed an encoder/decoder-based model that introduced Whitening and Coloring Transformations (WCT), achieving arbitrary stylization without the need for training specific styles. Li et al. [39] found that algorithms like AdaIN and WCT employed linear transformations between content image features and a transformation matrix for style transfer. They proposed a general optimization approach for linear style transfer, where a linear feedforward model was trained to replace matrix computations in the stylization process, simplifying the feedforward stylization model's generation process and improving generation speed. Shen et al. [40] introduced a novel meta-network for arbitrary style transfer. Compared to previous models, this meta-network had a smaller model size, making it more portable and suitable for running on mobile devices. Park et al. [8] discovered that previous algorithms did not balance global and local style patterns well. Therefore, they introduced the Style-Attentional Network (SANet), which flexibly matched style descriptors to content descriptors based on the semantic spatial distribution of the content image.

In summary, the flexibility of feedforward stylization models has greatly improved, enabling arbitrary style transfer. However, the generated image quality still lags slightly behind slow style transfer algorithms based on image iteration.

b: STYLIZATION METHODS BASED ON GAN NETWORKS

In 2014, GANs were introduced by Goodfellow et al. [41]. GANs consist of two components: a generator and a discriminator, as illustrated in Figure 2.

The training process can be seen as a game between the generator and the discriminator. The generator aims to produce realistic-looking data, while the discriminator's goal is to distinguish real data from fake data. Through adversarial training, both components learn and improve together to achieve the best possible generation results. Li et al. [42] used adversarial training to train a feedforward network based on MRF, resulting in realistic images. Their algorithm outperformed Johnson et al.'s [27] feedforward generation model by preserving coherent textures in complex images. However, Li et al. 's algorithm did not consider semantic relevance and performed less effectively on non-texture styles like facial synthesis. Mirza et al. [43] introduced Conditional GANs (CGANs) for image generation, extending the original GAN model by adding extra information to both the generator and the discriminator to guide the generation process. However, this supervised learning algorithm required training on perfectly matched pairs of data, which are not always available. This led to further research into unsupervised stylization using GANs. Zhu et al. [44] introduced CycleGAN, a GAN that does not require training on paired datasets. CycleGAN employs two generators and two discriminators. The generators are responsible for transforming images between two domains, while the discriminators differentiate between real and fake images in their respective domains. This model not only requires images to be transformed from the source domain to the target domain but also requires target domain images to be transformable back to the source domain, referred to as cycle-consistency loss. DisCoGAN [45] and DualGAN [46] share similarities with CycleGAN in terms of model architecture and experimental approaches. Liu et al. [47] designed the UNIT framework, which combines GANs and Variational Autoencoders (VAEs) to achieve unsupervised image-to-image translation by constructing a shared latent space. These generative adversarial training models addressed the limitation of CGANs, which required training on paired datasets. However, they could only learn the relationships between two different domains at a time, making it challenging to handle transformations between multiple domains. Choi et al. [48] introduced StarGAN, a model trained on multiple cross-domain datasets, enabling multi-domain transformations. StarGAN takes target domain labels as inputs to the generator, allowing the generator to produce different outputs based on varying target domain labels. It trains the discriminator to identify real or fake images and classify them into the appropriate domain. During style transfer, StarGAN only alters the domain-specific differences. Subsequently, Chen et al. [49] introduced CartoonGAN, which transforms real images into cartoon style. This model builds upon CycleGAN by adding two losses based on cartoon image features, i.e., edge adversarial loss and content loss, resulting in generated images with clear edges and content features resembling cartoon images. Pęśko et al. [50] extended CartoonGAN to transform videos into cartoon style by extracting keyframes. Li et al. [51] introduced the Attentive Adversarial Network (AAN) for cartoonizing selfies, known as SCGAN (Selfie Cartoonization Generative Adversarial Network). Wang et al. [52] improved CartoonGAN with a novel lightweight GAN known as AnimeGAN, which preserves the original colors of images and only cartoonizes textures. This is achieved through the introduction of three novel loss functions, i.e., grayscale style loss, color reconstruction loss, and grayscale adversarial loss. The resulting images in the style of Hayao Miyazaki are shown in Figure 3.

FIGURE 3. AnimeGAN stylization result. (a) Content image. (b) Miyazaki style image.

Following this, Zhao et al. [53] proposed ACL-GAN (Adversarial Consistency Loss-GAN), which encourages generated images to retain essential features of the source images rather than requiring complete translation back to the source domain. This allows images from two domains to focus on feature-level similarities rather than pixel-level similarities, enhancing flexibility and functionality. However, the above models only focused on transforming image styles while neglecting geometric differences between real and cartoon images. Therefore, Cao et al. [54] introduced CariGAN, which simultaneously performs stylization and geometric transformation on facial photos, resulting in images with both the texture style of cartoons and exaggerated geometric appearances. The core of geometric deformation lies in the introduction of a new loss function called feature loss. This loss exaggerates the most significant features by calculating the difference between facial coordinates in the input image and the average facial coordinates. Shi et al. [55] introduced WarpGAN, which combines CNNs and GANs to achieve fully automated transformation of facial photos into a cartoon style.

In summary, the use of GANs has provided new insights into the stylization field, improving both speed and image quality. Researchers have also designed various GANs tailored to specific requirements, enhancing the practicality of style transfer technology and facilitating its application in commercial settings.

III. THE PROPOSED METHOD

A. OVERALL ARCHITECTURE

In order to achieve content preservation and semantic-region style coherence in images, this paper proposes a novel architecture, which includes an encoder-decoder structure and CSAM, as shown in Figure 4. In this work, a pre-trained VGG19 network is employed as the encoder to extract deep features from both the content and style images, which are then input into CSAM. Within CSAM, the dual-stream content-style processing block (DS-CSPB) is used to perform channel-wise feature processing for style descriptors and position-wise feature processing for content descriptors, content-style matching attention block (CS-MAB) aligns the statistics of content descriptors with the attention-weighted mean and variance of style descriptors, the content-style

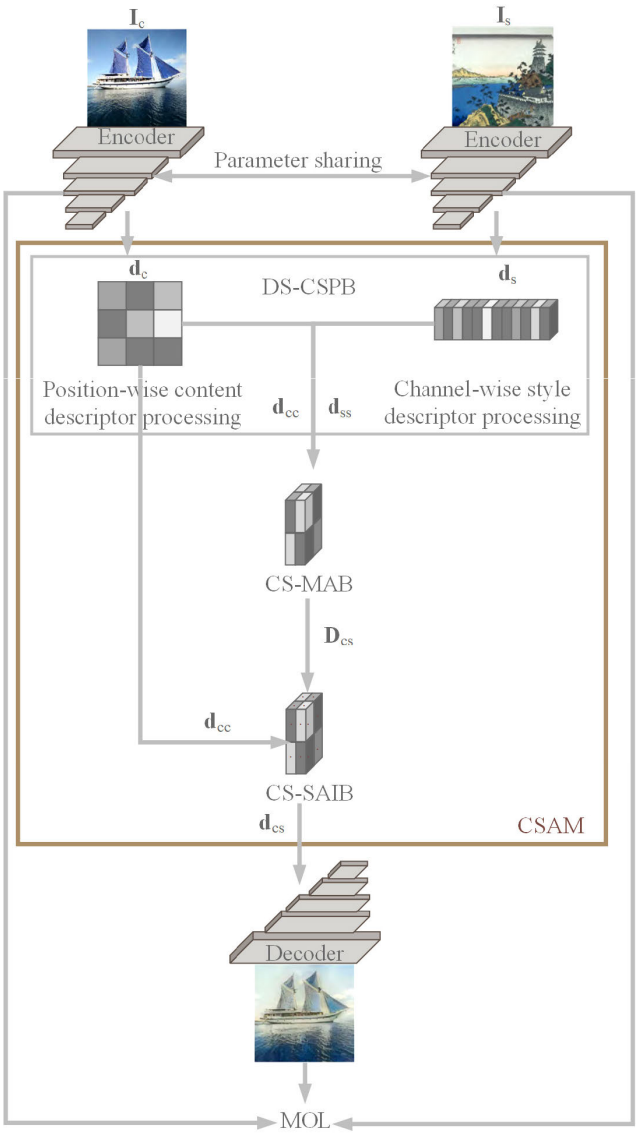

FIGURE 4. Overall architecture.

space-aware interpolation block (CS-SAIB) adaptively interpolates between the corresponding content and stylized descriptors to improve their feature matching degree. Finally, the enhanced features after fusion are passed through the decoder to generate stylized images, with the decoder structure being symmetric to the encoder structure.

B. CSAM

To achieve style coherence of content semantics while preserving the semantic structure of content images, this paper introduces a novel module designed for artistic image stylization, called CSAM. It consists of DS-CSPB, CS-MAB, and CS-SAIB, as shown in Figure 5.

1) DS-CSPB

This paper uses two attention-based streams to process the input content and style descriptors separately, aiming to

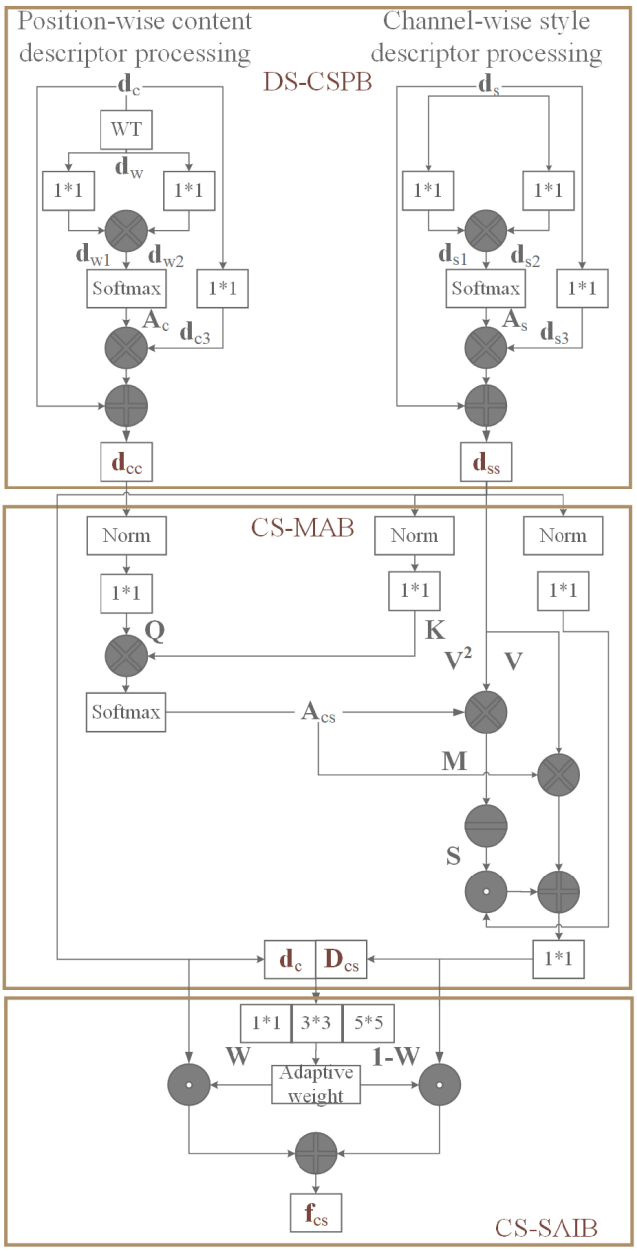

FIGURE 5. CSAM.

calculate their self-similarity through the attention mechanism for enhancing content and style representations. Firstly, since style descriptors can be represented using the inner product of vectorized feature maps, channel-wise feature processing operations are introduced to enhance the artistic styles in style images. The extracted style descriptors $\mathbf{d}_s \in \mathbf{R}^{C\times H\times W}$ are provided to three convolutional layers to generate $\mathbf{d}_{s1}$, $\mathbf{d}_{s2}$, and $\mathbf{d}_{s3}$, and reshaped to $\mathbf{R}^{C\times N}$, where $N = H\times W$. Then, the style attention map $\mathbf{A}_s \in \mathbf{R}^{C\times C}$ is formulated as (1):

$$\mathbf{A}_s = \text{softmax}(\mathbf{d}_{s1} \otimes \mathbf{d}_{s2}^T) \tag{1}$$

In which, $\otimes$ denotes the matrix-wise multiplication between feature maps $\mathbf{d}_{s1}$ and $\mathbf{d}_{s2}$. Finally, the enhanced style feature map is calculated as (2):

$$\mathbf{d}_{ss} = \mathbf{A}_s^T \otimes \mathbf{d}_{s3} + \mathbf{d}_s \tag{2}$$

To remove style-related texture information from content images while preserving their global structures, this paper transforms content descriptors $\mathbf{d}_c \in \mathbf{R}^{C\times H\times W}$ to generate feature map $\mathbf{d}_w$ through whitening transformation (WT).

As preserving local semantics of content images in stylized results is crucial, position-wise feature processing is introduced to adaptively capture detailed information in content descriptors. The newly processed feature maps $\mathbf{d}_{w1}$, $\mathbf{d}_{w2}$, and $\mathbf{d}_{c3}$ are reshaped to $\mathbf{R}^{C\times N}$ through convolutional operations, generating the content attention map $\mathbf{A}_c \in \mathbf{R}^{N\times N}$ as in (3) and the enhanced content feature map as in (4):

$$\mathbf{A}_c = \text{softmax}(\mathbf{d}_{w1}^T \otimes \mathbf{d}_{w2}) \tag{3}$$

$$\mathbf{d}_{cc} = \mathbf{d}_{c3} \otimes \mathbf{A}_c^T + \mathbf{d}_c \tag{4}$$

2) CS-MAB

Currently, most attention-based research methods re-weight the style feature maps and simply fuse them into content feature maps, which may cause misalignment between transferred style descriptors and corresponding content descriptors, leading to inappropriate colors and textures in the stylized images. To address this issue, this paper introduces a CS-MAB. This block learns the correspondence between content descriptors and style descriptors based on the attention weight information of content and style descriptors. It enables a more accurate embedding of local style patterns from the stylization references into content feature maps at each position. The block processes content feature maps $\mathbf{d}_{cc}$ and style feature maps $\mathbf{d}_{ss}$ through channel-wise mean-variance normalization, resulting in attention weight map $\mathbf{A}_{cs}$.

To align style descriptors with corresponding content descriptors more effectively in the feature space and ensure that content semantics are transferred appropriately to the style information, the proposed block re-weighted the mean value and standard variance of style descriptors, and expressed the weighted mean $\mathbf{M}\in\mathbf{R}^{C\times N}$ and the weighted standard variance $\mathbf{S}\in\mathbf{R}^{C\times N}$ as follows in (5) and (6):

$$\mathbf{M} = \mathbf{V} \otimes \mathbf{A}_{cs}^T \tag{5}$$

$$\mathbf{S} = \sqrt{(\mathbf{V}\odot\mathbf{V}) \otimes \mathbf{A}_{cs}^T - \mathbf{M}\odot\mathbf{M}} \tag{6}$$

In which, $\odot$ denotes element-wise multiplication. Finally, the transformed feature map is obtained by using the weighted standard deviation $\mathbf{S}$ and weighted mean $\mathbf{M}$:

$$\mathbf{D}_{cs} = \mathbf{S}\odot\text{Norm}(\mathbf{d}_{cc}) + \mathbf{M} \tag{7}$$

3) CS-SAIB

In the research of image stylization, most existing methods are primarily focused on how to transfer the artistic styles from the reference image to the content image as much as

possible, while neglecting the correlation between content semantics and a single style pattern. This often leads to multiple different artistic styles being transferred into the same content semantics, resulting in a visually confusing effect in the generated images. To address these issues, this paper introduces a space-aware interpolation block [56] to achieve a one-to-one matching effect between the semantic distribution of the content image and the style pattern of the style image. The space-aware interpolation block enables two types of descriptors to be aligned in the same space, allowing different semantic distributions in the content image and different style patterns in the style image to be matched separately. This block utilizes regional information for adaptive interpolation between $\mathbf{d}_{cc}$ and $\mathbf{D}_{cs}$, summarizing multi-scale regional information using three different scales of convolutional kernels, as in (8):

$$\mathbf{W} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Conv}_i(\mathrm{Concen}(\mathbf{d}_{cc}, \mathbf{D}_{cs})) \tag{8}$$

In which, $\mathrm{Conv}_i(.)$ represents the i-th convolutional kernel, Concen(.) denotes the channel-wise concatenation operation, and n is set to 3. The concatenated features help identify differences between corresponding content and style descriptors and address the local incoherence introduced by the previous attention block. The spatial weights $\mathbf{W} \in \mathbf{R}^{H \times W}$ obtained from the learnable channel-wise concatenation operation are used for interpolation:

$$\mathbf{d}_{cs} = \mathbf{W} \odot \mathbf{d}_{cc} + (1 - W) \odot \mathbf{d}_{cs} \tag{9}$$

Unlike previous feature fusion methods, CS-SAIB fuses features in the same spatial domain. Therefore, the interpolated content descriptors do not degrade, preserving the integrity of the content semantics. Finally, the stylized descriptors $\mathbf{d}_{cs}$ are fed into the decoder to generate the final stylized image.

C. MOL

The proposed multifaceted optimization loss function is denoted as L_{MOL}, which includes three main components: reconstruction loss (L_{rec}), content loss (L_c), and style loss (L_s), the latter of which consists of three sub-losses, namely L_{rEMD} (relaxed earth mover distance loss, rEMD), L_{MM} (moment matching loss, MM), and L_{CH} (color histogram loss, CH). The total loss function is formulated as in (10):

$$L_{MOL} = \alpha L_c + \beta L_{rec} + \underbrace{\gamma L_{rEMD} + \delta L_{MM} + \varepsilon L_{CH}}_{L_s} \tag{10}$$

In which, α, β, γ, δ and ε represent the weights assigned to L_c, L_{rec}, L_{rEMD}, L_{MM}, and L_{CH}, respectively, with values set to 2.0, 2.0, 0.5, 9.0, and 1.0.

1) rEMD loss

To effectively enhance the distribution of colors and textures from the style image onto the content image, the rEMD loss is used to optimize style descriptors during feature alignment, as in (11):

$$\begin{cases} L_{rEMD} = \max\left(\frac{1}{H_s W_s}\sum_i \min_j \mathbf{M}_{ij}, \frac{1}{H_s W_s}\sum_j \min_i \mathbf{M}_{ij}\right) \\ \mathbf{M}_{ij} = 1 - \frac{\mathbf{d}_{cs}^i \otimes \mathbf{d}_s^j}{||\mathbf{d}_{cs}^i|| \cdot ||\mathbf{d}_s^j||} \end{cases} \tag{11}$$

where i and j are the row order and column order of the matrix, respectively, $\mathbf{M}_{ij}$ represents a pairwise cosine distance matrix between $\mathbf{d}_{cs}$ and $\mathbf{d}_s$.

2) MM LOSS

To address issues with the cosine distance used in L_{rEMD}, which neglects the magnitude of feature vectors and may lead to visual artifacts, the MM Loss is introduced, as in (12):

$$L_{MM} = ||\mathbf{M}_{cs} - \mathbf{M}_s||_1 + ||\Sigma_{cs} - \Sigma_s||_1 \tag{12}$$

In which, $\mathbf{M}$ and Σ are the mean and covariance matrix of feature vectors.

3) CH LOSS

While the proposed module can generate high-quality stylized images, it also tends to produce color blending artifacts. Moreover, it causes color transfer to be overly uniform in certain regions of the content image and mixes various style patterns together. To address this issue, this paper refers to the differentiable color histogram loss introduced by AFIFI et al. [64]. This CH loss, at the cost of sacrificing partial coherence, effectively reduces the color mixing problem, as in (13):

$$L_{CH} = \frac{1}{2^{1/2}}||\mathbf{H}_s^{1/2} - \mathbf{H}_{cs}^{1/2}||_2 \tag{13}$$

where $\mathbf{H}$ is the color histogram feature.

4) CONTENT LOSS

In order to preserve the semantic structure of the content image and prevent significant deformations and distortions during the stylization process, this paper adopts the content loss proposed in [56]. This loss is based on the structural self-similarity between the content descriptor $\mathbf{d}_c$ and the stylized descriptor $\mathbf{d}_{cs}$, which is described as follows:

$$L_c = \frac{1}{H_c W_c}\sum_{i,j} \left|\frac{\mathbf{M}_{ij}^c}{\sum_i \mathbf{M}_{ij}^c} - \frac{\mathbf{M}_{ij}^{cs}}{\sum_j \mathbf{M}_{ij}^{cs}}\right| \tag{14}$$

In which, $\mathbf{M}_{ij}^c$ and $\mathbf{M}_{ij}^{cs}$ represent the pairwise cosine distance matrix between $\mathbf{d}_c$ and $\mathbf{d}_{cs}$.

5) RECONSTRUCTION LOSS

The existing attention transformation method alters both the original style feature space and the content feature space. This characteristic is detrimental to the learning process of the proposed CS-MAB and exacerbates issues related to

incoherent feature alignment. To address this challenge and constrain all features within the VGG space, we employ an image reconstruction loss. This loss compels the decoder to reconstruct VGG features, ensuring that all features between the encoder and decoder remain within the VGG space. This strategy helps maintain a shared space between content and style descriptors, as in (15):

$$L_{rec} = \mathrm{c}(||\mathbf{I}_{rc} - \mathbf{I}_c||_2 + ||\mathbf{I}_{rs} - \mathbf{I}_s||_2) + \sum_i (||\mathbf{R}_i(\mathbf{I}_{rc}) - \mathbf{R}_i(\mathbf{I}_c)||_2 + ||\mathbf{R}_i(\mathbf{I}_{rs}) - \mathbf{R}_i(\mathbf{I}_s)||_2) \tag{15}$$

In which, $\mathbf{I}_c$ and $\mathbf{I}_s$ are the input content and style images, $\mathbf{I}_{rc}$ and $\mathbf{I}_{rs}$ are the content and style images of the VGG feature reconstructions, c is a constant weight (set to 25), and $\mathbf{R}_i(\mathbf{I})$ represents the ReLU-i layer VGG features of image $\mathbf{I}$.

IV. EXPERIMENT

A. DATASET

In this paper, we utilized the MS-COCO dataset [57] for content and the WIKIART dataset [58] for benchmark style. The MS-COCO dataset comprises 82,783 natural photos across various categories worldwide, while the WIKIART dataset includes 80,095 artistic images from 27 different styles, serving as the training set. Furthermore, to showcase the method's generality and for comparison purposes, we adopted the settings from [65] and randomly selected 2000 images from the Places365 dataset [66] as an additional content dataset. For testing, we used 1,000 real photos and artistic images.

B. EVALUATION METRICS

Our evaluation about image stylization primarily considers qualitative evaluation and quantitative evaluation.

Qualitative evaluation focuses on visual consistency (i.e., whether the style aligns the content, including color and texture), visual quality (i.e., whether the image is clear and detailed), artistic effects (i.e., whether the image is attractive and aesthetically pleasing), visual authenticity (i.e., whether the image appears natural without unrealistic artifacts or distortions), and whether it meets users' aesthetic and emotional preferences (i.e., subjective evaluation criteria).

Quantitative evaluation methods in the MS-COCO content dataset primarily include Fréchet Inception Distance (FID), Content Fidelity (CF), Global Effects (GE), and Local Pattern (LP).

1) FID

FID is a metric used to compare the similarity between two data distributions, typically used to evaluate the performance of generative models. In image stylization, it is used to compare the distribution of style-transferred (ST) images with that of ground-truth (GT) images. A lower FID value indicates that the visual quality of the ST images is closer to that of GT images. Its calculation is as follows (16):

$$FID(ST, GT) = ||\mu_{ST} - \mu_{GT}||^2 + \mathrm{Tr}(\Sigma_{ST} + \Sigma_{GT} + 2(\Sigma_{ST}\Sigma_{GT})^{0.5}) \tag{16}$$

where GT∈{c, s}, representing content images and style images, respectively. μ is the mean of images, Σ is the covariance matrix of images, and Tr(.) denotes the trace.

2) CF

CF uses deep convolutional neural networks to extract high-level semantic features and calculates the similarity between the ST images and content images through the computation of multi-scale feature similarity.

3) GE

GE initially directly compares global color histograms obtained from style images and ST images, and then calculates multi-layer features in style images and ST images using the Gram matrix to better evaluate overall texture across multiple layers.

4) LP

LP extracts a set of 3 × 3 neural patches from multi-layer features of style images and ST images and calculates local style pattern similarity and diversity using normalized cross-correlation. Its calculation is as follows (17):

$$LP = \Sigma(1 - NCC(p_s, p_{ST})) \tag{17}$$

where p_s represents local feature patches extracted from style images, and p_{ST} represents local feature patches extracted from ST images. NCC denotes the normalized cross correlation as in (18):

$$NCC(p_s, p_{ST}) = \frac{\Sigma(p_s - \mu_{p_s})(p_{ST} - \mu_{p_{ST}})}{\sqrt{\Sigma(p_s - \mu_{p_s})^2(p_{ST} - \mu_{p_{ST}})^2}} \tag{18}$$

Note that CF, GE, and LP all use cosine similarity for comparison since cosine similarity is independent of feature dimensions and feature point values, making it better for measuring overall alignment in feature direction rather than absolute value differences.

Quantitative evaluation methods in the Places365 content dataset primarily include Content Loss [38] (CL), LPIPS [67], and Deception Rate (DR) [68]. For CL and LPIPS, we used a pre-trained VGG-19 and computed the average perceptual distance between the content images and stylized images. For DR, we trained a VGG-19 network for classifying 10 different styles from WikiArt. DR is then calculated as the percentage of stylized images for which the pre-trained network predicts the correct target style.

C. EXPERIMENT SETTINGS

Adam is employed as the optimizer with a learning rate of 0.0001. During training, 8 content-style image pairs are used in each batch. The input content and style images are resized

to 512×512 and then randomly cropped to 256×256 patches for effective training. The experiments are implemented using PyTorch 1.8.1 and trained in parallel on 4 NVIDIA Tesla V100 SXM2 GPUs with 32GB. The training process consists of 160,000 iterations with parameter checkpoints saved every 10,000 iterations. During testing, our architecture can handle images of arbitrary sizes.

D. BENCHMARK METHODS

To demonstrate the effectiveness of the proposed method, contrast experiments were conducted against various benchmark methods, including Style-Attentional Networks (SANet) [8], multi-adaptation network (MANet) [59], AdaAttN [10], and Progressive Attentional Manifold Alignment (PAMA) [56]. The following provides an introduction to these benchmark methods:

1) SANet

SANet employs attention mechanisms to selectively extract and apply style descriptors from reference images, capturing both global and local style patterns. It incorporates cross-feature style fusion to enhance style diversity. Through comprehensive experiments and visual examples, SANet showcases its ability to generate high-quality stylized images while preserving content details. This method provides a strong foundation for achieving impressive stylization results, making it a noteworthy benchmark for comparison with other style transfer methods.

2) MANet

MANet learns to adaptively match and blend content and style descriptors from a given image and a reference style image. This process allows for precise control over the degree of stylization while maintaining content structure. MANet incorporates multiple adaptation modules to capture varying levels of detail and style complexity, resulting in impressive style transfer results. This method serves as a valuable benchmark for comparing with other style transfer techniques, particularly in its ability to handle diverse style inputs.

3) AdaAttN

The core idea of AdaAttN involves adapting attention modules to selectively transfer style information, effectively preserving content details and achieving more coherent and visually appealing stylized images. AdaAttN's innovative attention mechanisms enable it to surpass conventional stylization methods, making it a compelling choice for comparative analysis in style transfer research.

4) PAMA

PAMA focuses on dynamically rearranging style descriptors based on the spatial distribution of content descriptors through attention operations. This approach enables the alignment of content and style manifolds on feature maps. This core idea of PAMA is crucial in improving the fidelity and coherence of style transfer, making it an essential point of reference for evaluating stylization methods.

TABLE 1. Quantitative assessment of image stylization effectiveness.

Methods /Metrics	SANet	MANet	AdaAttN	PAMA	Proposal 1
FID(ST, c)	483.95	499.84	357.16	483.99	300.04
FID(ST, s)	532.29	497.21	508.73	498.73	506.83
CF	0.51	0.47	0.53	0.50	0.57
GE	0.84	0.80	0.85	0.86	0.82
LP	0.50	0.48	0.49	0.50	0.52

TABLE 2. Quantitative assessment of image stylization efficiency about different sizes (MS).

Methods /sizes	SANet	MANet	AdaAttN	PAMA	Proposal
(256,256)	4.79	8.20	19.76	8.53	9.42
(512,512)	6.35	8.99	22.52	9.87	10.23

TABLE 3. Results of additional generality evaluation experiments.

Methods /Metrics	SANet	MANet	AdaAttN	PAMA	Proposal
CL	0.128	0.133	0.132	0.133	0.121
LPIPS	0.325	0.318	0.313	0.311	0.289
DR	0.544	0.510	0.517	0.593	0.650

E. CONTRAST EXPERIMENTS

1) QUANTITATIVE EVALUATION

a: EFFECTIVENESS CONTRAST

Table 1 provides a quantitative assessment of image stylization effectiveness.

We use 10 content images and 10 style images to generate 100 stylized images for each method and shows the average scores in Table 1. According to the FID metric, the proposed method achieves optimal results in preserving the structural semantics of content images while maintaining the competitive stylization compared to recent state-of-the-art methods. It also scores highest in CF and LP metrics, indicating a better balance between transferring artistic styles in local details and preserving overall content structure. Although it slightly lags in the GE metric compared to other methods, the sacrifice of some global style to ensure the overall visual quality of stylized images is acceptable.

b: EFFICIENCY CONTRAST

Table 2 compares the average stylization speeds of our method with other methods.

To ensure fairness, all methods are implemented using PyTorch. Experiments were conducted on NVIDIA Tesla V100 SXM2 GPUs for 256 px and 512 px images. The values in the table represent the average runtime for 100 image pairs. Since the proposed method is an improved approach based on attention mechanisms, it has the slightly slower speeds compared to SANet, but it achieves speeds similar to MANet

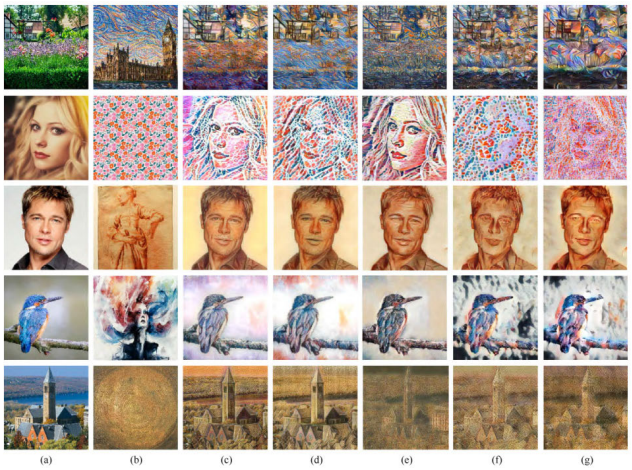

FIGURE 6. Comparative stylization results. (a) Content image. (b) Style image. (c) Proposal. (d) PAMA. (e) AdaAttN. (f) MANet. (g) SANet.

and PAMA, and is more than twice as fast as AdaAttN. In fact, the speed of our method is limited by attention blocks and could be further improved by designing a lightweight network in the future.

2) QUALITATIVE EVALUATION

a: OBJECTIVE EVALUATION

To evaluate the proposed method sensorally, we compare it with four other stylization methods, as shown in Figure 6.

In Figures 6(f) and 6(g), SANet and MANet utilize attention mechanisms to perform deep transformations of deep features. They calculate attention maps from both style and content descriptors and adjust their style descriptors, integrating attention outputs into content descriptors. While these methods generate fine-grained results with vivid style styles, including texture and color, they still suffer from issues such as semantic distortion (1st and 2nd rows) and visual artifacts (3rd and 5th rows), causing significant distortions in the global semantic structure of the content image. Since these attention-based methods independently present feature points without considering semantic distribution, they do not strictly match a single style style (4th row).

In Figure 6(e), AdaAttN extends the attention mechanism by fusing shallow features onto deep features, achieving better content semantic preservation. However, the increasing preservation of content semantics sacrifices style styles, leading to significant differences in color distribution compared to style images (2nd and 4th rows). It also introduces style confusion (1st row) and visual artifacts (3rd and 5th rows).

In Figure 6(d), PAMA aligns content descriptors with style descriptors through three attention alignment modules, gradually fusing style information into content descriptors, allowing the attention mechanism to capture feature distribution and maintaining global semantic structure well (4th and 5th rows). However, there are still local distortions (2nd and 3rd rows) and style mixing (1st row), resulting in confusing stylization effects.

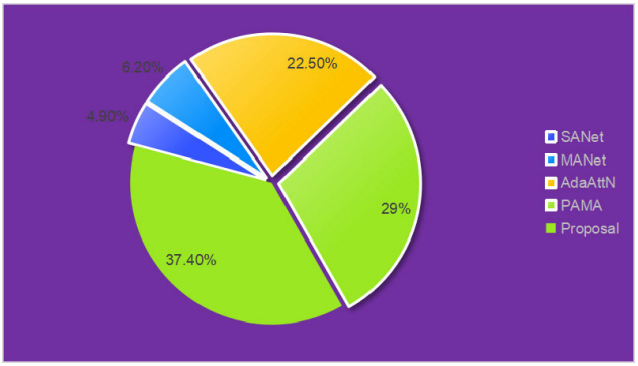

FIGURE 7. Subjective evaluation results.

Figure 6(c) represents our method, which processes deep content descriptors through whitening transformations, preserving the global structure and fine details of the content image effectively. It achieves content-style coherence by utilizing CS-MAB and CS-SAIB for one-to-one alignment of content semantics with artistic styles.

b: SUBJECTIVE EVALUATION

To provide statistical data evaluation of method performance, we conduct a user study with 100 participants to compare the visual effects of stylized images. To ensure the fairness and comprehensiveness of the user study, 40 participants were professionals in the field of image stylization, 30 were researchers in image processing-related fields, and 30 were individuals unrelated to image research. we randomly select 30 different content images and 30 different style images from the test set and paired them randomly. Each participant is required to select the image with the best artistic visual effect from the stylized images generated by the proposed method and the four comparison methods. A total of 3,000 votes are collected, and the voting ratios for each method are statistically analyzed, as shown in Figure 7.

From Figure 7, it can be seen that our method received the most votes, demonstrating that it not only has good content preservation capabilities but also generates stylized images with high style consistency and overall artistic visual effects.

F. ABLATION EXPERIMENTS

1) LOSS ABLATION

We provide results from loss ablation experiments to validate the effectiveness of each loss function used to train the proposed architecture as shown in Figure 8.

To test the effectiveness of the content loss L_c, it was removed, resulting in significant distortion and warping of the semantic structure of the content image, as shown in Figure 8(c).

Removing the reconstruction loss L_{rec} led to incoherent alignment between content and style descriptors, as shown in Figure 8(d).

Eliminating L_{rEMD} resulted in minimal style transfer, as shown in Figure 8(e).

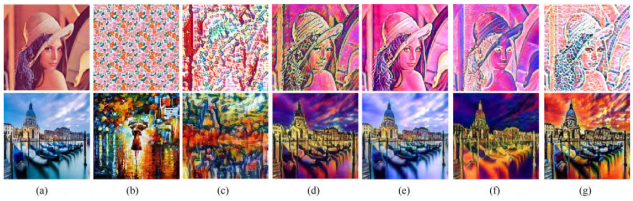

FIGURE 8. Loss ablation results. (a) Content image. (b) Style image. (c) without L_c. (d) without L_{rec}. (e) without L_{rEMD}. (f) without L_{CH}. (g) MOL.

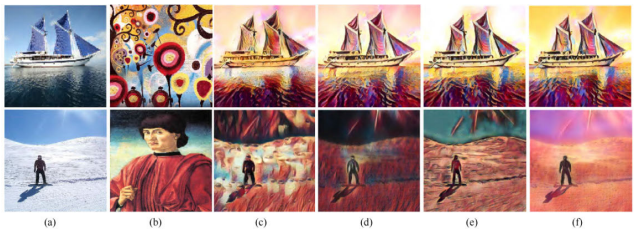

FIGURE 9. Block ablation results. (a) Content image. (b) Style image. (c) without WT. (d) without CS-SAIB. (e) without CS-MAB. (f) CSAM.

Removing L_{CH} caused color blending and overly uniform color transfer in a region of the content image, mixing various artistic styles, as shown in Figure 8(f).

Since removing L_{MM} would only generate the original content image, experiments without loss ablation were not performed.

These experiments confirm the effectiveness of the five selected loss functions, which are all indispensable for achieving the desired results in this paper.

2) BLOCK ABLATION

The proposed CSAM consists of three main blocks. To validate the necessity of all blocks for the experimental results, ablation experiments were conducted, as shown in Figure 9.

The DS-CSPB primarily relies on whitening transformations to normalize content descriptors, removing style information while preserving the global structure.

When whitening transformations are absent, too much of the original style information from the content image is retained, resulting in noticeable local structural distortion, as shown in Figure 9(c).

When CS-SAIB is removed, the features fused by CS-MAB cannot distinguish the differences between corresponding content and style descriptors. It also fails to identify the local incoherence caused by attention block, ultimately leading to the transfer of multiple artistic styles in content semantics and generating chaotic stylized images, as shown in Figure 9(d).

Omitting CS-MAB means replacing it with SANet. It can be observed that without the our fusion block, style descriptors are not accurately embedded at each position in the content feature map, leading to misalignments between content local semantics and inappropriate style information, as shown in Figure 9(e).

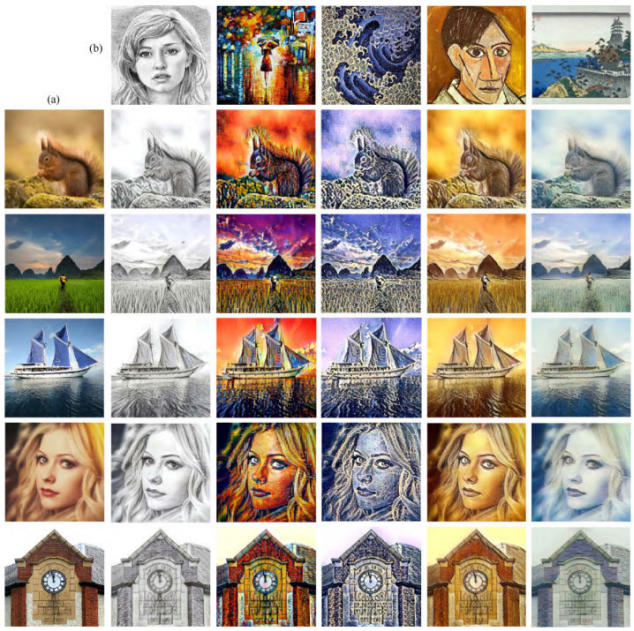

FIGURE 10. Visualizations. (a) Content image. (b) Style image.

G. VISUALIZATIONS

To visually evaluate the performance of the proposed method, we present some generated results of image stylization.

As shown in Figure 10, we select five different categories of source images, including animals, fields, oceans, buildings, and portraits, and five different categories of reference images, including sketches, oil paintings, crayon drawings, impressionist paintings, and landscapes, for image style transfer experiments. From the generated 25 stylized images, it can be observed that our method can perform arbitrary stylization experiments between various types of images. It not only preserves the global structure of the content image but also maintains the integrity of local features. Furthermore, it effectively conveys style information (i.e., texture and color) and achieves precise segmentation of different semantic structures, aligning perfectly with the artistic style. Ultimately, it produces visually pleasing stylized image results.

H. GENERALITY EXPERIMENT

To validate the universality of our method, a quantitative evaluation was performed on the additional Places365 dataset. The results are shown in Table 3.

It can be observed that even when using an additional content dataset, our method performs well in stylization, delivering highly competitive results.

V. DISCUSSIONS

While image stylization techniques have made significant research progress, there are still some areas where research is not sufficiently deep, and certain unresolved issues persist.

The following describes these problems and suggests some targeted improvement methods:

A. LACK OF STANDARDIZED EVALUATION METRICS FOR GENERATED IMAGE QUALITY

Currently, there is a lack of standardized objective evaluation systems for the quality of images generated by style transfer models. Evaluations often involve human judgments, which are highly subjective and lack scientific rigor. Therefore, a future research direction is to establish a standardized evaluation process and metric system. This could involve specifying an algorithm as a comparison standard, defining a set of standardized image datasets for evaluation, and involving a diverse range of evaluators, including both the general public and relevant artists, while providing them with a fixed evaluation framework.

B. INSUFFICIENT RESEARCH ON FAST SEMANTIC STYLE TRANSFER IN GENERATED MODELS

Current semantic segmentation is primarily applied to slow-style transfer models, with most researchers employing the VGG model for image feature extraction. While VGG is effective in extracting image features, it comes with significant computational complexity. Hence, a critical research direction is to enhance the generation quality of fast-style transfer models while reducing style leakage when incorporating semantic style transfer. A potential approach is to employ a feedforward stylization network for style transfer and then construct an image semantic stylization network to segment the input content image, identifying regions that require stylization. Finally, image fusion and edge smoothing can be performed on these regions.

C. TEXTURE-ONLY STYLE TRANSFER

Existing neural style transfer algorithms often transfer both color and texture simultaneously. However, there are scenarios where maintaining the color of the content image while applying style only to its texture is desired. Thus, achieving a higher degree of selectivity in generating images while preserving the color underlies a future development trend. This can be potentially realized by working with grayscale images, initially converting both the content and style images to grayscale. Then, only the texture features are transferred from the style image to the content image. Finally, a color transfer algorithm can be used to reconstruct the color of the stylized image based on the content image, ensuring both texture and color preservation.

D. PERSONALIZED PROCESSING

To enhance image effects and cater to specific domain requirements, it is possible to further investigate the incorporation of additional processing during the style transfer process. Introducing color transfer, as proposed by Zhang et al. [60], can facilitate color control in stylized images. The utilization of image fusion, as demonstrated by Luan et al. [61], can harmonize foreground and background images during stylization. Additionally, incorporating image segmentation for multimodal stylization, as introduced by Zhang et al. [62], allows for the transfer of different styles to segmented modules. Integrating these image processing techniques into stylization has significant implications for commercial applications.

E. CONSTRAINTS ON SHAPE ALTERATION

Most current stylizations primarily focus on altering image texture and color while neglecting the impact on image shape. In specific contexts, there is a need to generate images with shapes that resemble the target image more closely. For example, when converting real faces into cartoon characters, the transformation involves not only stylistic changes but also exaggerating the shape characteristics, such as the outline. Therefore, integrating geometric transformations with image stylization represents a crucial avenue for advancing neural style transfer models. This could involve training a deformation network that combines with the style transfer network to ensure that the input image closely matches the target image in terms of both style and shape, catering to artistic domains like cartoon production and filmmaking.

F. AUTO-TUNING

To achieve desirable stylized image results, manual parameter tuning is often required, particularly in model optimization-based methods. Each adjustment of model parameters usually necessitates retraining the model. While the prior method proposed a method for arbitrary stylization that does not require extensive training, alleviating the parameter tuning problem and avoiding the need to train separate models for different styles, the training process of this method is complex, and the image synthesis results are not significantly improved. Therefore, finding a simple, controllable, and quality-assured solution for auto-tuning parameters should be the focus of future research.

In summary, enhancing evaluation systems, improving model generation speed and quality, and increasing model flexibility and diversity to meet various commercial demands are the future research directions in the field of image stylization.

VI. CONCLUSION

In this work, we harness the power of AI to advance digital creativity design through image stylization. We propose a novel CSAM comprising DS-CSPB, CS-MAB, and CS-SAIB, effectively addressing the content-style discrepancy. In addition, our method leverages a MOL to optimize style and content descriptors, resulting in improved color and texture distribution while minimizing visual artifacts. This innovative method yields high-quality stylized images without significant content deformation.

REFERENCES

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015, *arXiv:1508.06576*.

[2] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1033–1038.

[3] L.-Y. Wei and M. Levoy, "Fast texture synthesis using tree-structured vector quantization," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 2000, pp. 479–488.

[4] M. Ashikhmin, "Synthesizing natural textures," in *Proc. Symp. Interact. 3D Graph.*, Mar. 2001, pp. 217–226.

[5] L. Liang, C. Liu, Y.-Q. Xu, B. Guo, and H.-Y. Shum, "Real-time texture synthesis by patch-based sampling," *ACM Trans. Graph.*, vol. 20, no. 3, pp. 127–150, Jul. 2001.

[6] C. Han, E. Risser, R. Ramamoorthi, and E. Grinspun, "Multiscale texture synthesis," in *Proc. ACM SIGGRAPH Papers*, Aug. 2008, pp. 1–8.

[7] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.

[8] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5873–5881.

[9] Y. Ma, C. Zhao, A. Basu, and X. Li, "RAST: Restorable arbitrary style transfer via multi-restoration," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 331–340.

[10] S. Liu, T. Lin, D. He, F. Li, M. Wang, X. Li, Z. Sun, Q. Li, and E. Ding, "AdaAttN: Revisit attention mechanism in arbitrary neural style transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6629–6638.

[11] A. Hertzmann et al., "Image analogies," in *Seminal Graphics Papers: Pushing the Boundaries*, vol. 2, 2023, pp. 557–570.

[12] B. Gooch, E. Reinhard, and A. Gooch, "Human facial illustrations: Creation and psychophysical evaluation," *ACM Trans. Graph.*, vol. 23, no. 1, pp. 27–44, Jan. 2004.

[13] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[14] G. Berger and R. Memisevic, "Incorporating long-range consistency in CNN-based texture generation," 2016, *arXiv:1606.01286*.

[15] E. Risser, P. Wilmot, and C. Barnes, "Stable and controllable neural texture synthesis and style transfer using histogram losses," 2017, *arXiv:1701.08893*.

[16] S. Li, X. Xu, L. Nie, and T.-S. Chua, "Laplacian-steered neural style transfer," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1716–1724.

[17] C. Castillo, S. De, X. Han, B. Singh, A. K. Yadav, and T. Goldstein, "Son of Zorn's lemma: Targeted style transfer using instance-aware semantic segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1348–1352.

[18] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6997–7005.

[19] S. Penhouët and P. Sanzenbacher, "Automated deep photo style transfer," 2019, *arXiv:1901.03915*.

[20] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," 2017, *arXiv:1701.01036*.

[21] C. Li and M. Wand, "Combining Markov random fields and convolutional neural networks for image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2479–2486.

[22] A. J. Champandard, "Semantic style transfer and turning two-bit doodles into fine artworks," 2016, *arXiv:1603.01768*.

[23] Y. L. Chen and C. T. Hsu, "Towards deep style transfer: A content-aware perspective," in *Proc. BMVC*, 2016.

[24] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. 2018, pp. 768–783.

[25] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang, "Visual attribute transfer through deep image analogy," 2017, *arXiv:1705.01088*.

[26] S. Gu, C. Chen, J. Liao, and L. Yuan, "Arbitrary style transfer with deep feature reshuffle," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8222–8231.

[27] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands. Springer, Oct. 2016, pp. 694–711.

[28] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," 2016, *arXiv:1603.03417*.

[29] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[30] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4105–4113.

[31] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," 2016, *arXiv:1610.07629*.

[32] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "StyleBank: An explicit representation for neural image style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2770–2779.

[33] K. Zhang, B. Wang, H.-S. Chen, X. Lei, Y. Wang, and C.-C. J. Kuo, "Dynamic texture synthesis by incorporating long-range spatial and temporal correlations," in *Proc. Int. Symp. Signals, Circuits Syst. (ISSCS)*, Jul. 2021, pp. 1–4.

[34] H. Zhang and K. Dana, "Multi-style generative network for real-time transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018.

[35] T. Q. Chen and M. Schmidt, "Fast patch-based style transfer of arbitrary style," 2016, *arXiv:1612.04337*.

[36] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.

[37] M. Lu, H. Zhao, A. Yao, F. Xu, Y. Chen, and L. Zhang, "Decoder network over lightweight reconstructed feature for fast semantic style transfer," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2488–2496.

[38] Y. Li et al., "Universal style transfer via feature transforms," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[39] X. Li, S. Liu, J. Kautz, and M.-H. Yang, "Learning linear transformations for fast arbitrary style transfer," 2018, *arXiv:1808.04537*.

[40] F. Shen, S. Yan, and G. Zeng, "Neural style transfer via meta networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8061–8069.

[41] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[42] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands. Springer, Oct. 2016, pp. 702–716.

[43] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *Comput. Sci.*, pp. 2672–2680, 2014, doi: 10.48550/arXiv.1411.1784.

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[45] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1857–1865.

[46] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2868–2876.

[47] M. Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[48] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[49] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "CartoonGAN: Generative adversarial networks for photo cartoonization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9465–9474.

[50] M. Pęśko, A. Svystun, P. Andruszkiewicz, P. Rokita, and T. Trzciński, "Comixify: Transform video into comics," *Fundamenta Informaticae*, vol. 168, nos. 2–4, pp. 311–333, Sep. 2019.

[51] X. Li, W. Zhang, T. Shen, and T. Mei, "Everyone is a cartoonist: Selfie cartoonization with attentive adversarial networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 652–657.

[52] X. Wang and J. Yu, "Learning to Cartoonize using white-box cartoon representations supplementary materials," Tech. Rep.

[53] Y. Zhao, R. Wu, and H. Dong, "Unpaired image-to-image translation using adversarial consistency loss," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Springer, Aug. 2020, pp. 800–815.

[54] K. Cao, J. Liao, and L. Yuan, "CariGANs: Unpaired photo-to-caricature translation," 2018, *arXiv:1811.00222*.
[55] Y. Shi, D. Deb, and A. K. Jain, "WarpGAN: Automatic caricature generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10754–10763.
[56] X. Luo, Z. Han, L. Yang, and L. Zhang, "Consistent style transfer," 2022, *arXiv:2201.02233*.
[57] T. Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland. Springer, Sep. 2014, pp. 740–755.
[58] F. Phillips and B. Mackintosh, "A case for critical thinking," *Issues Accounting Educ.*, vol. 26, no. 3, pp. 593–608, 2011.
[59] Y. Deng, F. Tang, W. Dong, W. Sun, F. Huang, and C. Xu, "Arbitrary style transfer via multi-adaptation network," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2719–2727.
[60] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands. Springer, Oct. 2016, pp. 649–666.
[61] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep painterly harmonization," *Comput. Graph. Forum*, vol. 37, no. 4, pp. 95–106, Jul. 2018.
[62] Y. Zhang, C. Fang, Y. Wang, Z. Wang, Z. Lin, Y. Fu, and J. Yang, "Multimodal style transfer via graph cuts," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5942–5950.
[63] A. Mordvintsev, C. Olah, and M. Tyka. (2015). *Inceptionism: Going Deeper Into Neural Networks*. [Online]. Available: https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html
[64] M. Afifi, M. A. Brubaker, and M. S. Brown, "HistoGAN: Controlling colors of GAN-generated and real images via color histograms," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7937–7946.
[65] Y. Zhang, F. Tang, W. Dong, H. Huang, C. Ma, T.-Y. Lee, and C. Xu, "A unified arbitrary style transfer framework via adaptive contrastive learning," *ACM Trans. Graph.*, vol. 42, no. 5, pp. 1–16, Oct. 2023.
[66] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
[67] H. Chen et al., "Artistic style transfer with internal-external learning and contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 26561–26573.
[68] A. Sanakoyeu, D. Kotovenko, S. Lang, and B. Ommer, "A style-aware content loss for real-time HD style transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 698–714.

LANTING YU was born in Wuhan, Hubei, China, in 1987. She received the master's degree from Chongqing University, China. She is currently with the School of Publishing and Media, Chongqing Business Vocational College, Chongqing. Her research interests include image stylization and new media art.

QIANG ZHENG was born in Yubei, Chongqing, China, in 1985. He received the master's degree from Southwest University, China. He is currently with CISDI Information Technology Company Ltd. His research interests include intelligent optimization algorithm, intelligent interaction technology, database technology, industrial internet, and data governance.