Efficient Hi-Fi Style Transfer via Statistical Attention and Modulation

Zhirui Fang[1], **Yi Li**[1*], **Xin Xie**[1], **Chengyan Li**[1], **Yanqing Guo**[1]
[1]Dalian University of Technology
{fangzr, shelsin, gggmdgzsx}@mail.dlut.edu.cn, {liyi, guoyq}@dlut.edu.cn

Abstract

Style transfer is a challenging task in computer vision, aiming to blend the stylistic features of one image with the content of another while preserving the content details. Traditional methods often face challenges in terms of computational efficiency and fine-grained content preservation. In this paper, we propose a novel feature modulation mechanism based on parameterized normalization, where the modulation parameters for content and style features are learned using a dual convolution network (BiConv). These parameters adjust the mean and standard deviation of the features, improving both the stability and quality of the style transfer process. To achieve fast inference, we introduce an efficient acceleration technique by leveraging a row and column weighted attention matrix. In addition, we incorporate a contrastive learning scheme to align the local features of the content and the stylized images, improving the fidelity of the generated output. Experimental results demonstrate that our method significantly improves the inference speed and the quality of style transfer while preserving content details, outperforming existing approaches based on both convolution and diffusion.

1 Introduction

The phrase "There are a thousand Hamlets in a thousand people's eyes" vividly encapsulates the complexity of stylized customization. Style transfer aims to seamlessly blend the content of a given image with the style of a target image while preserving the structural integrity and details of the original content. However, style is inherently a multidimensional concept, encompassing intricate information that is difficult to precisely articulate through language. These subtle stylistic details pose a challenge for traditional manual methods, making it particularly difficult to describe and reproduce styles with precision. As a result, style transfer remains a formidable task, requiring an approach that captures the diversity of styles while ensuring faithful style representation without compromising content consistency.

Traditional style transfer methods often struggle to balance content preservation and style fidelity. Early approaches, primarily optimization-based techniques, offer flexibility but are computationally expensive and suffer from slow convergence. More recent learning-based methods, leveraging deep neural networks, have significantly improved efficiency. However, they continue to face challenges in effectively modulating style features while maintaining fine-grained content structures. A key issue is the difficulty in precisely aligning style and content representations, often leading to artifacts, structural distortions, or inadequate style adaptation.

To address these challenges, researchers have explored various techniques, including adaptive instance normalization [Huang and Belongie, 2017], attention mechanisms[Yao *et al.*, 2019], and feature transformation strategies[Li *et al.*, 2017]. However, these methods still face limitations in capturing spatial dependencies and local style variations. Recently, several approaches have made significant progress. For instance, AdaAttN [Liu *et al.*, 2021] improves style-content alignment by applying adaptive normalization at each spatial location, leveraging both shallow and deep features to reduce local distortions. Inversion-Based Style Transfer [Zhang *et al.*, 2023] employs image inversion techniques to better capture and transfer artistic styles, improving both style fidelity and content preservation. ArtBank [Zhang *et al.*, 2024] employs a pre-trained diffusion model guided by an implicit style prompt bank to generate high-fidelity stylized images while preserving content structure, leveraging spatial-statistical attention to enhance training efficiency and style controllability.

However, despite recent advancements in style transfer, existing methods still face challenges in achieving both computational efficiency and high-quality results. Maintaining fine-grained content structure while accurately modulating style remains a significant obstacle. In most cases, the trade-off between style adaptation and content preservation limits performance. Furthermore, many methods struggle with efficient inference, especially in deeper networks, hindering real-time style transfer applications. As shown in Figure 1, while some methods can produce visually appealing results, they often fail to preserve content details or adapt to complex style patterns, highlighting the gap between quality and efficiency in current approaches.

*Corresponding author.

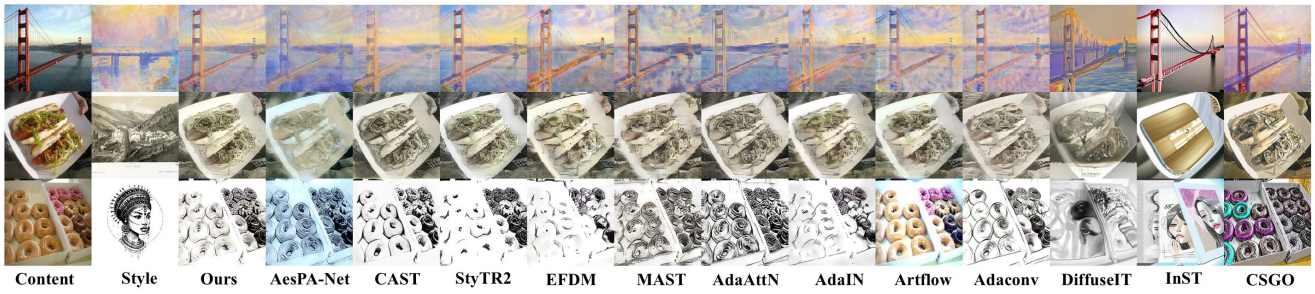

Figure 1: Stylization results. Given input images (columns 1-2), we compare our method (column 3) with both convolutional models (columns 4-12) and diffusion models (columns 13-15).

Inspired by the growing need for efficient and high-quality style transfer, we propose a novel approach for high-fidelity artistic style transfer with Statistical Row-Column Attention and Statistical Modulation (SRCA-SM). Our method addresses several key challenges in traditional style transfer techniques, such as maintaining content consistency while adapting to diverse stylistic variations, by designing two core components: 1) Statistical Row-Column Attention (SRCA), a mechanism that accelerates model inference speed by applying row and column weighted attention to modulate the interaction between content and style features. and 2) Statistical Modulation, which uses learned statistical parameters, including mean and standard deviation, to modulate both content and style features, enhancing style fidelity while preserving fine-grained content details. Furthermore, 3) Contrastive Learning is integrated to improves the fidelity of the style transfer by promoting the alignment of local features between the content and the stylized image. It leverages a contrastive loss to encourage similar features to be closer in the feature space, while distinguishing them from features at different spatial positions.

Our method is inspired by the observation that both content and style representations play a crucial role in determining the quality of the output, especially when they are modulated effectively. By decoupling content and style representations in a statistical manner, we achieve superior style transfer results that not only preserve intricate content details but also faithfully apply the desired stylistic transformations. Our contributions are as follows:

- We propose the Statistical Row-Column Attention (SRCA) mechanism, which accelerates model inference by applying row- and column-weighted attention to modulate the interaction between content and style features.
- We develop the Statistical Modulation (SM) that adjusts content and style features using learned statistical parameters, enhancing the quality of the generated stylized images without sacrificing content fidelity.
- We integrate contrastive learning to refine the feature alignment, improving the consistency of local features and ensuring that the content and style features are well-aligned.
- Our method achieves superior style transfer results with both quality and stability, as evidenced by extensive experimental evaluations.

2 Related Work

Gatys et al. [Gatys *et al.*, 2016] pioneered the use of deep convolutional networks to successfully fuse the structural content of a source image with the artistic style of a target image, achieving high-quality image style transfer. This groundbreaking work laid the foundation for the rapid development of style transfer techniques. Following this, many researchers have built upon Gatys et al.'s approach, proposing various improvements and extensions. These advancements primarily focus on enhancing computational efficiency, improving image quality, and optimizing the fusion of style and content, further advancing the field of image style transfer.

Chen et al. [Chen *et al.*, 2021] proposed a novel internal-external style transfer method that integrates both internal and external learning, significantly bridging the gap between human-created and AI-generated artworks. This approach innovatively combines internal and external features, enhancing the fusion of content and style and improving the overall effectiveness of style transfer. Liu et al. [Liu *et al.*, 2021] introduced a new attention mechanism modulation technique to enhance the performance of arbitrary style transfer. Their AdaAttn method refines the way content and style features are attended to during the transfer process, allowing for finer control over the style and increasing the transfer's flexibility and accuracy. Further advancing the field, Deng et al. [Deng *et al.*, 2022] introduced the Transformer architecture into style transfer with their StyTr^2 method. By leveraging the global self-attention mechanism of Transformers, StyTr^2 effectively captures long-range dependencies between style and content images, overcoming the limitations of traditional convolutional networks in processing local features and significantly improving style transfer performance, especially for complex style transformations.Despite these advancements, existing methods still face challenges in terms of balancing content preservation and style fidelity, especially when dealing with complex or diverse styles. Additionally, some approaches struggle with computational efficiency, limiting their applicability in real-time or large-scale scenarios.

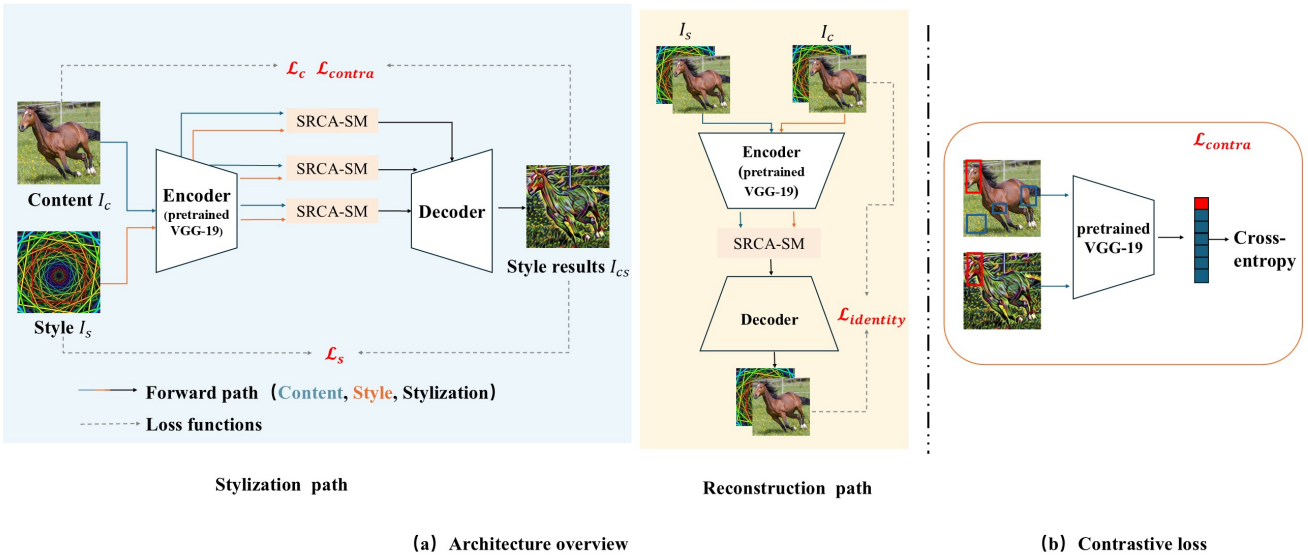

Figure 2: The overview of the Statistical Row-Column Attention and Statistical Modulation(SRCA-SM). (a) shows the architecture of SRCA-SM and training procedures, *i.e.*, stylization task (blue box) and reconstruction task (yellow box). (b) depicts the detail of contrastive loss.

Recently, Diffusion models (DMs) have achieved remarkable success in the field of image generation, thanks to their powerful generative capabilities, which enable the creation of high-quality, detailed, and diverse images. Diffusion models have demonstrated exceptional performance not only in generating high-resolution images but also in a variety of image generation tasks, including style transfer, super-resolution reconstruction, and image inpainting. Particularly in the area of style transfer, diffusion models achieve more refined style blending by carefully controlling the image generation process. Zhang et al. [Zhang *et al.*, 2023] introduced an inversion-based style transfer method, combining it with diffusion models. Unlike traditional style transfer methods, this approach maps the style image into a latent space and generates the style-transferred image through the diffusion model. Yang et al. [Yang *et al.*, 2023] proposed a text-guided diffusion image style transfer method based on zero-shot contrastive loss, combining diffusion models with text guidance. They designed a zero-shot contrastive loss function that allows the style transfer process to be guided by textual descriptions, eliminating the need for pre-trained image-style pairs or explicit style labels.Xing et al. [Xing *et al.*, 2024] introduced CSGO, a method for content-style composition in text-to-image generation, addressing the challenge of combining textual descriptions with specific content and style elements in image synthesis. Their approach focuses on seamlessly integrating both content and style, providing enhanced control over the generated images to ensure they align more closely with the desired output.Although diffusion models have achieved remarkable results in image generation and style transfer, they still suffer from high computational costs, especially when generating high-resolution images, leading to significant time and resource consumption. Furthermore, while they capture style details more effectively, challenges remain in maintaining stability and consistency when dealing with highly complex or diverse styles.

3 Proposed Method

Let I_c and I_s represent the content image and style image, respectively. The objective of our study is to extract the stylistic features of the style image by analyzing the differences in mean and variance between the content and style images and to transfer these features to the content image, thus generating high-quality stylized images. The proposed model workflow is illustrated in Figure 2.

Initially, both the content image and the style image are processed through the Statistical Row-Column Attention (SRCA) module to generate a preliminary realistic stylized image. Subsequently, the generated image undergoes a high-fidelity stylization process through modulation operations involving mean and variance. Finally, contrastive loss is used to further optimize the content fidelity of the generated image, ensuring semantic consistency with the original content image.

3.1 Statistical Row-Column Attention

In this section, we will provide a detailed explanation of the Statistical Row-Column Attention (SRCA) mechanism and present concrete proof that the smoothing mechanism of the attention matrix, weighted by row and column factors, significantly improves model efficiency by accelerating inference speed.

To compute the attention map A of layer x, we formulate Q (query), K (key) and V (value) as:

$$\begin{aligned} Q &= f(\text{Norm}(F_c^{1:x})) \\ K &= g(\text{Norm}(F_s^{1:x})) \\ V &= h(F_s^x) \end{aligned} \tag{1}$$

where f, g and h are learnable convolution layers, whereas Norm represents the normalizing features based on channel means and standard deviations. The attention map A can be calculated as:

$$A = Softmax(Q^T \otimes K) \tag{2}$$

where $\otimes$ denotes matrix multiplication. To improve the model efficiency and accelerate inference, we compute the mean of matrix A along its rows and columns, then integrate these results into the formulation:

$$\hat{A} = \alpha \cdot A \cdot W_{\text{col}} + (1-\alpha) \cdot A \cdot W_{\text{row}} \tag{3}$$

where $W_{\text{col}} \in R^{H_c W_c \times 1}$ and $W_{\text{row}} \in R^{1 \times H_c W_c}$ represent the column-wise and row-wise mean weights of A , respectively. The operator $\cdot$ denotes the element-wise product, and $\alpha \in (0, 1)$ is a hyperparameter that balances the contributions of the two weights. Futhermore,

$$M = V \cdot \hat{A} \tag{4}$$

Then, we can obtain the attention-weighted standard deviation $S \in R^{C \times R_c W_c}$ as:

$$S = \sqrt{(V \cdot V) \otimes \hat{A} - M \cdot M} \tag{5}$$

Finally, corresponding scale S and shift M are used to generate transformed feature map:

$$\hat{F^x_{cs}} = S \cdot Norm(F^x_c) + M \tag{6}$$

The target update function $L(W)$, where W represents the model parameters, is optimized using the gradient descent method. The update rule is expressed as: $w^{(t+1)} = w^{(t)} - \eta \nabla_w L$, where the gradient $\nabla_w L$ is influenced by both feature mapping and computation of the attention matrix. In our model, the design of the attention matrix A is a critical factor affecting the fluctuation of $\nabla_w L$, formulated as:

$$\nabla_W L = \frac{\partial L}{\partial \hat{F^x_{cs}}} \cdot G_{\hat{A}} \cdot G_{A,QK} \cdot G_W \tag{7}$$

where $G_S = \frac{\partial \hat{F^x_{cs}}}{\partial \hat{A}} \cdot \frac{\partial \hat{A}}{\partial A} + \frac{\partial \hat{F^x_{cs}}}{\partial M} \cdot \frac{\partial M}{\partial \hat{A}} \cdot \frac{\partial \hat{A}}{\partial A}$ represents the gradient contribution of S during the generation of the stylized features. $G_{A,QK} = \frac{\partial A}{\partial (Q,K)}$ captures the gradient flow through the attention matrix A. $G_W = \frac{\partial (Q,K)}{\partial W}$ characterizes the dependence of Q and K on the model parameters W.

Suppose that the elements of A are independent and identically distributed random variables with a mean and variance of μ_A and σ_A^2. Using the statistical formula for the variance of means, the variances of W_{col} and W_{row} are given by:

$$\text{Var}(W_{\text{col}}) = \frac{\sigma_A^2}{n_c}, \quad \text{Var}(W_{\text{row}}) = \frac{\sigma_A^2}{n_s} \tag{8}$$

where n_c and n_s represent the number of channels of content feature and the number of position of style feature, respectively. Incorporating the smoothing effects of W_{col} and W_{row} into the variance of $\hat{A}$, we obtain $\text{Var}(\hat{A}) \propto \alpha^2 \cdot \frac{\sigma_A^2}{n_c} + (1-\alpha)^2 \cdot \frac{\sigma_A^2}{n_s}$. Since the variance of gradient fluctuations is proportional to the variance of $\hat{A}$, we have:

$$\text{Var}(\nabla_w L) \propto \alpha^2 \cdot \frac{\sigma_A^2}{n_c} + (1-\alpha)^2 \cdot \frac{\sigma_A^2}{n_s} \tag{9}$$

This indicates that the mean operation applied to the row and column weights reduces the range or spread of the gradient fluctuations, effectively limiting the variance of the gradient by factors related to $\frac{1}{n_c}$ and $\frac{1}{n_s}$, thus significantly decreasing the noise in the gradient and improving the inference efficiency of the model.

3.2 Enhanced Style Transfer with Statistical Modulation

In the task of image style transfer, how to fine-tune the style features while maintaining content fidelity is a key issue. To address this challenge, we propose a style transfer method based on mean and variance modulation. This approach finely adjusts the features of both the content and the style images, achieving a high-quality transfer of style features.

Firstly, features are extracted from the x layers of the content image and concatenated to form a composite feature map. Subsequently, the concatenated content features, along with the preliminary stylized image, are fed into BiConvNet, a network composed of two convolution layers. The specific operation can be expressed as follows:

$$\begin{aligned} \alpha^c, \beta^c &= BiConv(F_c^{1:x}) \\ \alpha^{cs}_{std}, \beta^{cs}_{std} &= BiConv(\hat{F^x_{cs}}) \\ \alpha^{cs}_{mean}, \beta^{cs}_{mean} &= BiConv(\hat{F^x_{cs}}) \end{aligned} \tag{10}$$

Subsequently, we compute the mean and standard deviation for both the content and style images. This operation is carried out using the following formulas:

$$\begin{aligned} \mu_c, \sigma_c &= \mathcal{M}(F^x_c) \\ \mu_s, \sigma_s &= \mathcal{M}(F^x_s) \end{aligned} \tag{11}$$

$\mathcal{M}(\cdot)$ denotes the operation that computes the mean and standard deviation of the input data. Then, the mean and variance of the style image are adjusted using the adjustment parameters obtained in the convolution network:

$$\begin{aligned} \hat{F_c} &= Norm(F_c) * \alpha^c + \beta^c \\ \hat{\mu_s} &= \mu_c \cdot \alpha^{cs}_{mean} + \beta^{cs}_{mean} \\ \hat{\sigma_s} &= \sigma_s \cdot \alpha^{cs}_{std} + \beta^{cs}_{std} \end{aligned} \tag{12}$$

Finally, the stylized image is obtained through the equation $F_{cs} = \hat{F_c} \cdot \hat{\sigma_s} + \hat{\mu_s}$, where $\hat{F_c}$ is the content feature after modulation, with $\hat{\sigma_s}$ and $\hat{\mu_s}$ are the modulated standard deviation and mean of the style feature, respectively.

3.3 Loss Function

Internal style learning. It focuses on learning and preserving the inherent stylistic elements within the image, which are critical for maintaining the unique characteristics of the style while simultaneously adapting to content changes. This

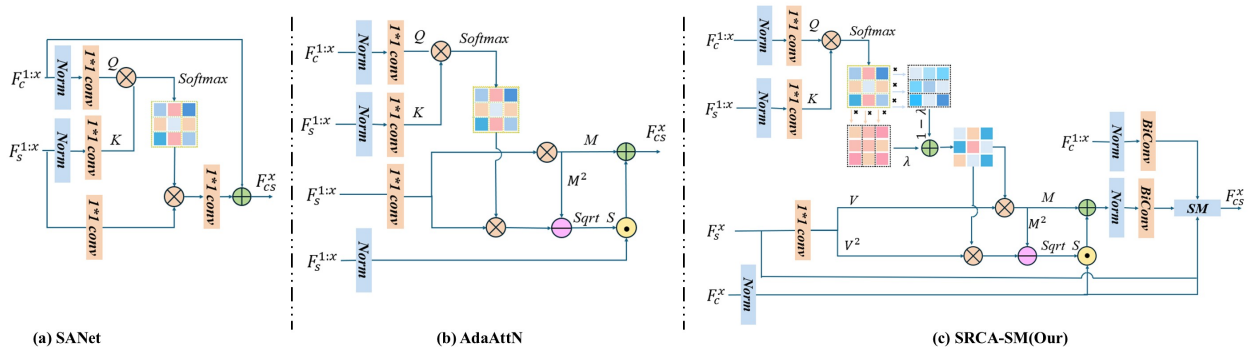

Figure 3: (a) The structure of SANet[Park and Lee, 2019]; (b) The structure of AdaAttN[Liu *et al.*, 2021]; ,(c) The structure of our proposed SRCA-SM.

process helps ensure that the stylistic features are effectively transferred from the target style to the content image. The loss function used for internal style learning is defined as:

$$\begin{aligned}\mathcal{L}_s = &\sum_{i-1}^{L} \|\mu(\phi_i(I_{cs})) - \mu(\phi_i(I_s))\|_2 \\ &+\|\sigma(\phi_i(I_{cs})) - \sigma(\phi_i(I_s))\|_2\end{aligned} \tag{13}$$

where ϕ denotes the feature extractor at the i-th layer, I_{cs} is the stylized image, and I_s is the target style image. μ and σ represent the mean and standard deviation of the extracted features, respectively.

Content structure preservation. It guarantees the preservation of the semantic and structural integrity of the original image, even as stylistic elements are modified, by minimizing the disparity in content features between the generated and target images.This can be formalized using the content loss function $\mathcal{L}_c$, which is typically defined as:

$$\begin{aligned}\mathcal{L}_c = &\|\phi_{conv4\_1}(I_{cs}) - \phi_{conv4\_1}(I_c)\| \\ &+\|\phi_{conv4\_4}(I_{cs}) - \phi_{conv4\_4}(I_c)\|\end{aligned} \tag{14}$$

Identity loss. Similar to [Park and Lee, 2019], we utilize the identity loss to learn richer and more accurate content and style representations. Specifically, the identity loss is designed to minimize the discrepancy between the input and the output when the content and style images are identical, as formulated below:

$$\begin{aligned}\mathcal{L}_{identity} = &\lambda_{identity1}(\|I_{cc} - I_c\|_2 + \|I_{ss} - I_s\|_2) \\ &+\lambda_{identity2}\sum_{i=1}^{L}(\|\phi_i(I_{cc} - \phi_i(I_c))\|)_2 \\ &+(\|\phi(I_{ss} - \phi(I_s))\|)_2\end{aligned} \tag{15}$$

Structural Information Preservation via Contrastive Learning. In [Chen *et al.*, 2021], it has been established that contrastive learning can achieve more satisfactory stylization results by capturing the relationships between stylized pairs. Motivated by this finding, we investigate whether contrastive learning can effectively preserve structural information by maximizing the mutual information between input and output patches. Specifically, the content image I_c and the stylized image I_{cs} are processed through a VGG network to extract their respective features, z and $\hat{z}$, which are then utilized for block-based contrastive loss computation. Pixels are randomly sampled from the feature maps for cross-entropy loss calculation. Pixels from identical spatial locations are designated as 'positives', with their mutual information maximized, whereas pixels from different spatial locations are treated as 'negatives', with their mutual information minimized. This process is formally defined as:

$$\mathcal{L}_{contra}(I_c, I_{cs}) = \mathbb{E}_{I_c}[\sum_{l}\sum_{b}\ell(z_l^s, \hat{z}_l^s, z_l^{B\setminus b})] \tag{16}$$

Here, z_l and $\hat{z}_l$ denote the l-th layer features from I_c and I_{cs}, respectively. The variable b represents a specific position sampled from B_l , where B_l refers to the set of all spatial positions in z_l. The term $B_l \setminus b$ denotes all positions in B_l except b , and the loss ℓ enforces the similarity between z_l^s and $\hat{z}l^s$ at position b while contrasting them with features at other positions in $B \setminus b$. This design encourages local feature alignment between I_c and I_{cs} , while maintaining distinctiveness across different spatial regions.

The entire network is optimized by minimizing the combined loss functions for style preservation ($\mathcal{L}_s$), content structure preservation ($\mathcal{L}_c$), identity loss ($\mathcal{L}_{identity}$), and structural information preservation via contrastive learning ($\mathcal{L}_{contra}$), and the overall optimization objective of the network can be expressed as the following combined loss function:

$$\mathcal{L} = \lambda_s\mathcal{L}_s + \lambda_c\mathcal{L}_c + \mathcal{L}_{identity} + \lambda_{contra}\mathcal{L}_{contra} \tag{17}$$

This ensures effective stylization while preserving the content structure and identity of the original image.

Figure 4: Qualitative comparison with convolutional models (4th-12th columns) and diffusion models (13th-15th columns).

4 Experiments

4.1 Implementing Details

We utilize the MS-COCO dataset [Phillips and Mackintosh, 2011] for content images and the WikiArt dataset [Phillips and Mackintosh, 2011] for style images. During the training phase, all images are randomly cropped to a fixed resolution of 256 × 256 pixels, whereas during testing, images of arbitrary resolution are supported. The model is optimized using the Adam optimizer [Kingma, 2014], with an initial learning rate of 0.0001 and a warm-up strategy for adjustment. The batch size is set to 8, and the network undergoes a total of 320,000 iterations during training.The loss function incorporates multiple terms, where the weights λ_s, λ_c,$\lambda_{identity1}$,$\lambda_{identity2}$ and λ_{contra} are set to 10, 8, 70, 1 and 0.1, respectively, ensuring a balanced yet flexible contribution from style loss, content loss, and contrastive loss.

4.2 Qualitative Comparisons

As shown in Figure 4, we compare our method with ten state-of-the-art arbitrary style transfer approaches, including traditional convolution-based methods such as AesPA-Net [Hong *et al.*, 2023], CAST [Zhang *et al.*, 2022b], StyTR^2 [Deng *et al.*, 2022], EFDM [Zhang *et al.*, 2022a], MAST [Deng *et al.*, 2020], AdaAttn [Kingma, 2014], Adain [Huang and Belongie, 2017], ArtFlow [An *et al.*, 2021] and AdaConv [Chandran *et al.*, 2021] , as well as diffusion-based methods such as DiffuseIT [Kwon and Ye, 2022], InST [Zhang *et al.*, 2023], and CSGO [Xing *et al.*, 2024]. To observe the model ability to tackle diverse styles, four different types of style images are presented, with whose characters are discussed in the following. We also zoom in all the results to present the details better.

Our method demonstrates superior performance in several challenging scenarios. In the first row, our approach excels in preserving fine details, such as the clarity of the cat's eyes, outperforming other methods that introduce distortions or artifacts. In the second row, our method effectively retains intricate textures, such as the clarity of the text on the "Abbey Road NW8" street sign, while eliminating artifacts and blurring that are present in other results. In the third row, our approach preserves critical content details, such as the human head's outline and fine features, while mitigating the cloud-like distortions and blurring present in other methods. Finally, in the fourth row, our method accurately transfers the black-and-white artistic style while preserving the fine details and structure of the original content image. Overall, these results highlight the ability of our approach to achieve high-fidelity style transfer by effectively balancing style fidelity, content preservation, and noise reduction, outperforming both traditional convolution-based and diffusion-based methods.

4.3 Quantitative Comparisons

Comparison with Conventional Transfer. As shown in Table 1, we evaluate the performance of our proposed method and compare it with several state-of-the-art style transfer techniques. The evaluation is conducted on a comprehensive dataset consisting of 20 content images and 40 style images, resulting in a total of 800 stylized images. This diverse dataset ensures that the reported metrics comprehensively reflect the robustness and adaptability of the methods.

Our method significantly outperforms conventional style transfer techniques in terms of ArtFID, achieving the lowest score of 30.244, which aligns well with human perceptual preferences. This indicates that our approach generates stylized images with a high degree of aesthetic appeal and a seamless blend of content and style. Furthermore, our method achieves competitive FID scores (18.905), demonstrating that the stylized images maintain a close resemblance to the target style distributions.

Metric	Ours	AesPA-Net	CAST	StyTR^2	EFDM	MAST	AdaAttn	AdaIN	ArfFlow	AdaConv	DiffIT	InST	CSGO
ArtFID↓	**30.244**	31.923	34.685	30.720	34.605	31.282	30.350	30.933	34.630	31.856	40.721	40.633	33.716
FID↓	18.905	19.764	20.395	18.890	20.062	**18.199**	18.658	18.242	21.252	19.022	23.065	21.571	19.197
LPIPS↓	**0.5194**	0.5448	0.6212	0.5445	0.6430	0.6293	0.5439	0.6076	0.5562	0.5910	0.6921	0.8002	0.6693
CSFD↓	**0.2452**	0.2499	0.2918	0.3011	0.3346	0.3043	0.2862	0.3155	0.2920	0.3600	0.3428	0.6759	0.4577
Time/sec↓	0.0934	0.5111	0.2917	0.2925	0.0962	0.0987	0.3575	**0.05375**	0.0790	0.2122	38.07	8.013	5.027

Table 1: Performance comparison of different methods on style transfer. Metrics include ArtFID, FID, LPIPS, CSFD, and Inference Time.

Method	Baseline	Improved baseline (+SM)	Content Align (+$\mathcal{L}_{contra}$)	Ours (+SRCA)
ArtFID↓	30.350	30.997	30.300	30.244
FID↓	18.658	18.553	18.989	18.905
LPIPS↓	0.6076	0.5852	0.5210	0.5194
CSFD↓	0.3155	0.2598	0.2510	0.2452
Time/sec ↓	0.3575	0.3780	0.3800	0.0934

Table 2: Ablation study results showing the performance of different variants on various metrics.

For content fidelity, our method achieves superior scores in both LPIPS (0.5194) and CSFD (0.2452). Notably, the significant reduction in CSFD highlights the ability of our method to preserve spatial correlations within the content image, even when applying challenging styles. These results validate the effectiveness of our approach in achieving high-quality style transfer while retaining fine-grained content details.

Comparison with Diffusion-based Style Transfer. Our method also demonstrates superior performance compared to diffusion-based style transfer techniques, achieving the best scores across LPIPS, FID, and ArtFID. As shown in Table 1, the proposed approach not only enhances perceptual quality but also ensures consistent and reliable style adaptation. While diffusion-based methods are often limited by their high computational cost and extended inference time due to iterative synthesis, our method achieves a significantly faster inference time of 0.0934 seconds per image, which is orders of magnitude faster than typical diffusion-based approaches.

This improvement in computational efficiency is particularly evident in our lightweight design and streamlined operations, making it suitable for both real-time applications and large-scale deployments. These advantages, combined with high fidelity and content consistency, validate the practicality of our method in real-world scenarios.

4.4 Ablation Studies

To evaluate the contribution of each component, we conducted an ablation study comparing the baseline method, the improved baseline, the model with self-supervised contrastive learning(+$\mathcal{L}_{contra}$) and final method (+ SRCA) across several performance metrics. In terms of image quality, the improved baseline showed a noticeable improvement over the original baseline with the introduction of SM. Adding self-supervised contrastive learning(+$\mathcal{L}_{contra}$) further enhanced style consistency and content preservation, especially in perceptual similarity (LPIPS) and content structure fidelity (CSFD). The final model (+ SRCA) demonstrated a significant improvement in both style consistency and content preservation, as SRCA effectively captured finer content-style relationships, leading to better overall image quality.

Regarding inference time, although other models showed increased computational costs, the inclusion of SRCA resulted in a significant reduction in inference time. This highlights the efficiency of the method, making it suitable for practical applications without compromising quality.

As shown in Table 2, the proposed components lead to consistent improvements across all metrics, confirming the effectiveness of each contribution.

5 Conclusion

In this paper, we propose a novel framework that effectively learns style information from a diverse set of style images and dynamically modulates the content structure and style patterns of input images. Our method generates highly realistic stylized images with harmonious style patterns and fine-grained textures, while preserving the structural integrity of the input content images. Extensive quantitative and qualitative experiments, conducted on a comprehensive dataset of 800 stylized images, demonstrate that our proposed SRCA-SM framework significantly outperforms state-of-the-art convolutional and diffusion-based methods in terms of ArtFID, LPIPS, CSFD, and computational efficiency.

Acknowledgments

This work is supported by the Major Program of the National Social Science Foundation of China under Grant No.19ZDA127.

References

[An *et al.*, 2021] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 862–871, 2021.

[Chandran *et al.*, 2021] Prashanth Chandran, Gaspard Zoss, Paulo Gotardo, Markus Gross, and Derek Bradley. Adaptive convolutions for structure-aware style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7972–7981, 2021.

[Chen *et al.*, 2021] Haibo Chen, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, Dongming Lu, et al. Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems*, 34:26561–26573, 2021.

[Deng *et al.*, 2020] Yingying Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. Arbitrary style transfer via multi-adaptation network. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2719–2727, 2020.

[Deng *et al.*, 2022] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11326–11336, 2022.

[Gatys *et al.*, 2016] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[Hong *et al.*, 2023] Kibeom Hong, Seogkyu Jeon, Junsoo Lee, Namhyuk Ahn, Kunhee Kim, Pilhyeon Lee, Daesik Kim, Youngjung Uh, and Hyeran Byun. Aespa-net: Aesthetic pattern-aware style transfer networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22758–22767, 2023.

[Huang and Belongie, 2017] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[Kingma, 2014] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kwon and Ye, 2022] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022.

[Li *et al.*, 2017] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017.

[Liu *et al.*, 2021] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6649–6658, 2021.

[Park and Lee, 2019] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5880–5888, 2019.

[Phillips and Mackintosh, 2011] Fred Phillips and Brandy Mackintosh. Wiki art gallery, inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3):593–608, 2011.

[Xing *et al.*, 2024] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024.

[Yang *et al.*, 2023] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. Zero-shot contrastive loss for text-guided diffusion image style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22873–22882, 2023.

[Yao *et al.*, 2019] Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. Attention-aware multi-stroke style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1467–1475, 2019.

[Zhang *et al.*, 2022a] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8035–8045, 2022.

[Zhang *et al.*, 2022b] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–8, 2022.

[Zhang *et al.*, 2023] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10156, 2023.

[Zhang *et al.*, 2024] Zhanjie Zhang, Quanwei Zhang, Wei Xing, Guangyuan Li, Lei Zhao, Jiakai Sun, Zehua Lan, Junsheng Luan, Yiling Huang, and Huaizhong Lin. Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7396–7404, 2024.