Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

StyleMamba: Efficient Image Style Transfer with Bidirectional Selective Scan Vision Mamba

Jian Liu
SICHUAN NORMAL UNIVERSITY
Jun Yang
jkxy_yjun@sicnu.edu.cn

SICHUAN NORMAL UNIVERSITY
DiWei Wu
SICHUAN NORMAL UNIVERSITY
Hewen Liu
SICHUAN NORMAL UNIVERSITY
Jun Liu
SICHUAN NORMAL UNIVERSITY

Research Article

Keywords: Image Style Transfer, Bidirectional Selective Scan Mamba, Enhance Efficiency

Posted Date: February 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8742918/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

StyleMamba: Efficient Image Style Transfer with Bidirectional Selective Scan Vision Mamba

Jian Liu[1], Jun Yang[2*], Diwei Wu[3], Hewen Liu[4], Jun Liu[5]

[1] Sichuan Normal University, Chengdu, Sichuan, China.

*Corresponding author(s). E-mail(s): jkxy_yjun@sicnu.edu.cn;
Contributing authors: jl@stu.sicnu.edu.cn; WDW525@stu.sicnu.edu.cn;
l526982483@163.com; 1164811474@qq.com;

Abstract

Style transfer aims to render stylized images with artistic features while maintaining the original content. Traditional CNN-based approaches have limitations in dealing with global information and long-range dependencies in style transfer. Existing transformer-based approaches, while mitigating these problems well, face high computational costs. Mamba addresses these limitations using a selective structured state-space model (S4), which maintains linear complexity while effectively handling long-range dependencies. In this pa-per, we propose StyleMamba, an efficient image style transfer architecture based on a Bidirectional Selective Scan mechanism, to address the challenge of balancing local and global dependencies with the computational efficiency of existing methods through spatial dyadic state space modeling. In addition, we design Dynamic Gate Fusion (DGF) to fuse dual-path outputs based on feature relevance adaptively. Performing qualitative and quantitative experiments on representative datasets, we demonstrate the advantages of our model over other state-of-the-art(SOTA) transformer-based methods and different approaches.

Keywords: Image Style Transfer, Bidirectional Selective Scan Mamba, Enhance Efficiency

1 Introduction

Image style transfer is a method for incorporating the artistic features of a stylized image into a content image. It has evolved from traditional style transfer methods [8] to style transfer methods performing iterative optimization [12] and feed-forward network-based style transfer methods [2]. On this basis, the fusion of multiple styles remains a challenging problem, which has been addressed by [14, 15, 19] using generic style transfer. However, these approaches are not sufficiently fine-grained in modeling content-style interactions, failing to achieve the desired results in some cases.

To solve this problem, StyTr2 [6] adopts the Transformer architecture to enhance the global modeling capability of style transfer, but the global attention mechanism of its encoder leads to the loss of local details. S2WAT [25] improves local feature extraction by strip-window attention mechanism, which effectively mitigates the local detail problem, but the introduced multi-window fusion strategy increases the computational complexity. PUFF-Net [28] adopts a single-encoder design and features a decoupling module to reduce the number of parameters, but it cannot maintain the local details well. Although Vmamba[17] reduces the computational overhead, its four-directional scanning mechanism may sometimes lead to the loss of finegrained content details in stylized images during the decoding process.

To address the above limitations, we propose StyleMamba an efficient image style migration framework based on a BSS mechanism (Bidirectional Selective Scan). Local-global collaborative modeling is achieved by the independence of rows and columns Mamba paths, and combined with a dynamic gated fusion strategy, enhances inference efficiency while improving the capability of detail preservation. Our approach involves the design of a bidirectional selective scan feature extractor based on the Mamba2 architecture for efficient modeling of multi-scale features through heterogeneous directional scan and dynamic feature

fusion. To capture comprehensive feature dependencies, a horizontal dynamic scan constructs long-range interactions along the image width via input-dependent state transitions. Concurrently, a vertical parity rearrangement strategy, which separates and reorders rows, strengthens local structural relationships. To achieve input-sensitive global-local feature equalization, we propose Dynamic Gate Fusion(DGF), which adaptively assigns dual-path contribution weights by light-weighting the feature relevance of the network's implicitly learned channel dimensions.

In summary, the main contributions of this paper can be summarized as follows:

- We propose a Bidirectional Selective Scan(BSS) framework framework based on the Mamba2 architecture for efficient modeling of multi-scale features through the synergy of horizontal dynamic parameterization paths and vertical structural reorganization paths. The design maintains content details while preserving the linear complexity of Mamba.
- To address information redundancy and conflict in dual-branch feature integration, we design a Dynamic Gate Fusion(DGF) module that achieves fine-grained feature calibration through learnable gating factors, enabling adaptive two-branch feature fusion.
- Quantitative and qualitative experiments demonstrate that, compared to mainstream baseline models, this approach achieves a synergistic enhancement in both content detail retention and stylistic quality whilst significantly improving inference efficiency.

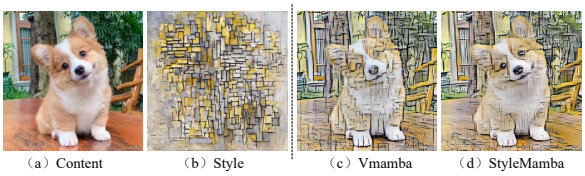


Fig. 1 Visual comparison using Vmamba and StyleMamba as encoders respectively

As evidenced by the visual comparisons in Figure 1, in a direct comparison where both methods utilize Vmamba as the encoder, our StyleMamba, however, achieves a superior balance by better preserving critical content details during the stylization process

2 Related Work

2.1 Image Style Transfer

[9, 22] proposed an image style migration method based on Convolutional Neural Networks (CNN), which captures the content structure and style texture information of an image by using different layers of the network respectively and iteratively generates a new type of image fusing the target content and style using an optimization algorithm. Single style transfer [23] uses an end-to-end model to achieve realtime style transfer for one specific style. Later extended applications of multi-style transfer [2] combine multiple styles in a single model to achieve the desired stylization. In recent years, arbitrary style transfer [18, 19, 27] has been based on generalization capabilities. ArtFlow [1] uses Projected Flow Networks (PFNs) to achieve unbiased content results, InST [26] achieves stylistic transfer via SDM and IEST [3], and CAST [27] generates excellent stylization results using contrast learning techniques. StyTr2 [6] successfully generates high-quality stylization results by using the transformer architecture as a core component. S2WAT [25] proposes a hierarchical visual transformer that dynamically fuses short-range and long-range dependencies through multi-shape window attention (horizontal/vertical strip-like and square windows) and adaptively assigns weights by using the "Attn Merge" strategy to solve the localization problem caused by traditional window attention. Puff-Net [28] designs a lightweight encoder-only transformer, introducing a pure content/style feature extractor to separate content structure and style texture, reducing computational costs and improving inference speed.

2.2 Mamba

State Space Models (SSMs) [10] have demonstrated formidable competitiveness in deep learning, rooted in classical control theory and enabling long-range dependency modeling through linear time complexity. Structured State Space Sequence Models (S4) and Mamba [17] are both based on a continuous-system discretization framework. Mamba2 [5] introduces Structured State Space Duality (SSD), unifying SSMs with attention mechanisms via semiseparable matrices while designing hardware-efficient algorithms. The mathematical formulation of Mamba-2's Structured State Space Duality is:

$$h_t = A_t h_{t-1} + B_t x_t, \quad y_t = C_t^\top h_t \tag{1}$$

where $A_t \in \mathbb{R}^{N\times N}$ is the dynamic state matrix, and $B_t \in \mathbb{R}^{N\times 1}$, $C_t \in \mathbb{R}^{N\times 1}$ are time-varying projection parameters. Compared to Mamba, Mamba2 simplifies the structure to improve hardware efficiency by constraining to A_t scalar unit matrices ($A_t = a_t I$) and optimizes the computational complexity to $O(TN^2)$ training FLOPs and $O(TN)$ inference FLOPs by block-decomposition matrix multiplication, which is 2-8 times faster than Mamba's scanning implementation. In the

discretization process, Mamba2 follows the zero-order holding rule, which transforms the continuous parameters $(\hat{A}, \hat{B})$ into discrete parameters:

$$A = \exp(\Delta\hat{A}), \quad B = (\Delta\hat{A})^{-1}(\exp(\Delta\hat{A}) - I)\Delta\hat{B} \tag{2}$$

While applications for Mamba2 are still emerging, its architectural advances in structured state-space duality (SSD) and hardware-aware efficiency provide a solid foundation for cross-domain extensions. In the meantime, the original Mamba architecture has demonstrated remarkable generality across different domains. In the computer vision direction, the framework has been successfully applied to scenarios such as medical image segmentation (e.g., the VM-Unet [21] model for adaptive feature extraction), video timeseries modeling, and multi-spectral image fusion; and in the field of graph data modeling, Mamba has enabled the learning of efficient graph structural representations through serialized node embedding [20]. Notably, the research community has spawned several innovative variants: FMamba [13] improves the quality of image defogging through multilevel feature fusion, while Mambadfuse [16] innovatively inte-grates residual networks with state-space modeling to achieve significant results in cross-modal target detection tasks [7]. Such advances validate the scalability potential of the Mamba architecture for complex visual representation tasks.

3 Method

This section presents an overview of the proposed StyleMamba's overall architectural design, as illustrated in Figure 2(d).

3.1 Bidirectional Selective Scan Mamba

3.1.1 Bidirectional Selective Scan

The shape of the input feature is $C \times H \times W$, where C is the number of channels, H is the height and W is the width.Scanning is done using two parts: row-selective scan and column-selective scan. The row-selective scan generates sequences by parity across rows, which are subjected to global context modeling by the SS2D module. The column selective scan is similar to the row-selective scan in that it generates sequences with parity across columns, and again the SS2D module performs locally-aware feature selection. Integrating these two sequences enables a balance between global and local connections. As shown in Figure 3, this fusion strategy achieves an effective balance between computational efficiency and integrated perceptual capabilities by pooling information from a wider region. The corresponding row-selective scan formula is:

$$x_{\text{row}} = \mathcal{S}_{\text{row}}(x) \in \mathbb{R}^{C\times\frac{H}{2}\times 2W} \tag{3}$$

$$\mathcal{F}_{\text{row}} = \text{SS2D}\left(\Gamma(x_{\text{row}})\right) \in \mathbb{R}^{C\times(H\cdot W)} \tag{4}$$

where $\mathcal{S}_{\text{row}}(\cdot)$ splits the input feature x into two sub-tensors $(\mathbb{R}^{C\times\frac{H}{2}\times W})$ by parity rows, which are spliced into $(\mathbb{R}^{C\times\frac{H}{2}\times 2W})$ along the width dimension to form a long procedural column covering the full width, which is serialized by $\Gamma(\cdot)$ and fed into the SS2D module for modeling global dependencies. Similarly, the corresponding column-selective scan formula is:

$$x_{\text{col}} = \mathcal{S}_{\text{col}}(x^{\top}) \in \mathbb{R}^{C\times\frac{W}{2}\times 2H} \tag{5}$$

$$\mathcal{F}_{\text{col}} = \text{SS2D}\left(\Gamma(x_{\text{col}})\right) \in \mathbb{R}^{C\times(H\cdot W)} \tag{6}$$

where $x^{\top}$ is the transpose of x, $\mathcal{S}_{\text{col}}(\cdot)$ is split into $(\mathbb{R}^{C\times\frac{W}{2}\times H})$ by parity and spliced into $(\mathbb{R}^{C\times\frac{W}{2}\times 2H})$ along the height dimension to form a vertically oriented fine-grained sequence, which is serialized by $\Gamma(\cdot)$ and fed into the SS2D module to model the local dependencies. This approach is able to capture local-global connections with a much lower number of parameters while maintaining optimal computational efficiency.

3.1.2 Dynamic Gate Fusion

The Dynamic Gate Fusion module implements feature fusion of bi-sequences in three steps: firstly, the bidirectional scan features are spliced along the channel dimensions; secondly, the inter-feature correlation matrix is built by a learnable linear transformation; finally, the gating matrix is decomposed into row/column weight components for feature fusion and combined with DropPath regularization, the fused features are residually linked to the initial input features Linking. Through the synergistic optimization of dynamic gating-guided bi-directional feature fusion and random path regularization, an adaptive balance between global semantics and local details is achieved, thus enhancing the cross-scale context modeling capability. The process is implemented as follows:

$$\mathsf{K} = Concat(\mathcal{F}_{row}, \mathcal{F}_{col}) \in \mathbb{R}^{N\times 2d} \tag{7}$$

$$G = Softmax(K\mathcal{W}_g) \in \mathbb{R}^{N\times 2d} \tag{8}$$

$$G_{row}, G_{col} = Split\,(\,G\,),\ G_{row}, G_{col} \in \mathbb{R}^{N\times d} \tag{9}$$

$$\mathcal{F}_{\text{fused}} = x + \text{DropPath}\big(G_{\text{row}} \odot \mathcal{F}_{\text{row}} + G_{\text{col}} \odot \mathcal{F}_{\text{col}}\big) \in \mathbb{R}^{N\times d} \tag{10}$$

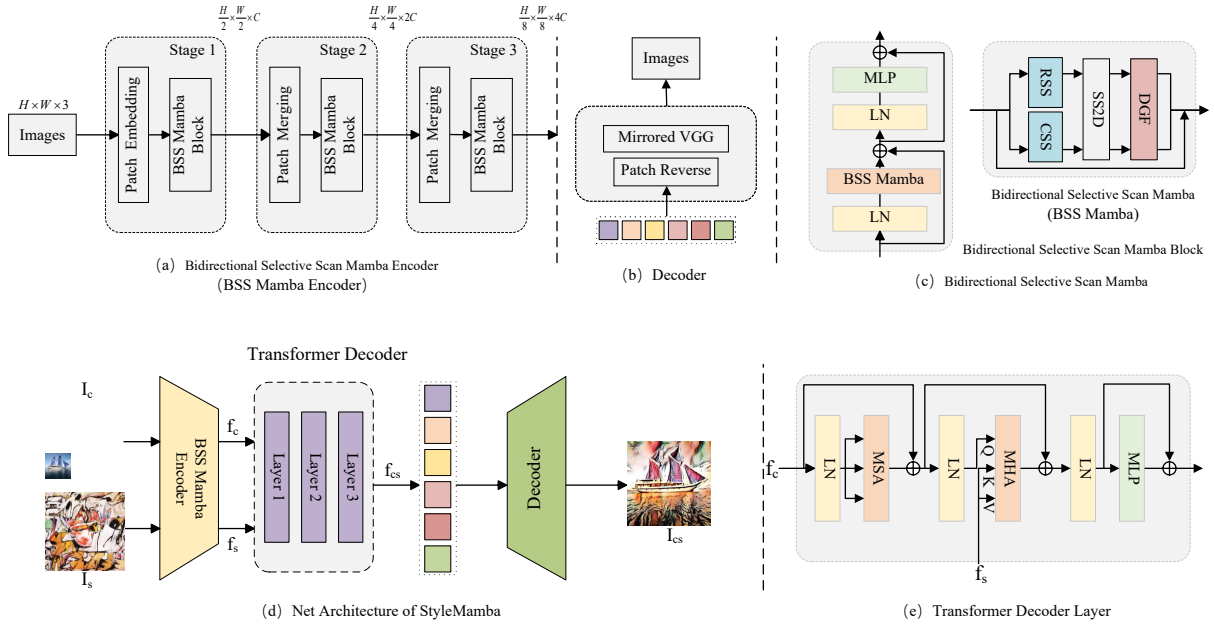


Fig. 2 The overall pipeline of the proposed StyleMamba. Encoder extracts content (f_c) and style (f_s) features from inputs I_c and I_s. The transfer module stylizes f_c with f_s to generate f_{cs}, which the decoder decodes into output I_{cs}.

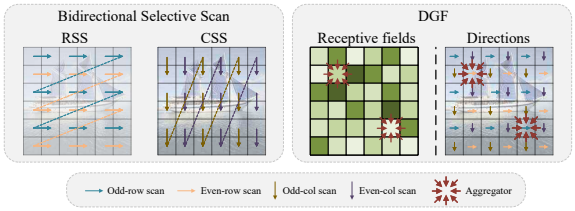


Fig. 3 Overview of the proposed method. The Bidirectional Selective Scan module generates two sequences from row-selective scan and column-selective scan, which are fused by our Dynamic Gate Fusion to achieve local-to-global linkage

3.1.3 Bidirectional Selective Scan Mamba Block

We now provide an overview of the workflow of the BSS Mamba block. The structure of a BSS Mamba block is essentially analogous to that of a standard Transformer block, with the exception that the conventional Multi-head Self Attention (MSA) module is substituted with a BSS Mamba module. As illustrated in Figure 2(a), a BSS Mamba block primarily comprises a module and an MLP module with GELU as anon linear activation function. Layer normalization (LN) operations are conducted prior to each module and MLP, while the outputs are integrated via residual connections following the completion of each module. The computational flow of a BSS Mamba block is as follows:

$$\mathcal{F}_{\text{row}} = \text{SS2D}(\Gamma(\mathcal{S}_{\text{row}}(\text{LN}(Z^{l-1})))) \tag{11}$$

$$\mathcal{F}_{\text{col}} = \text{SS2D}(\Gamma(\mathcal{S}_{\text{col}}(\text{LN}(Z^{l-1})^{\top}))) \tag{12}$$

$$\tilde{Z}^{l} = \text{DropPath}(\mathcal{G}(\mathcal{F}_{\text{row}}, \mathcal{F}_{\text{col}})) + Z^{l-1} \tag{13}$$

$$Z^{l} = \text{MLP}(\text{LN}(\tilde{Z}^{l})) + \tilde{Z}^{l} \tag{14}$$

where $\mathcal{G}(\cdot)$ denotes the 'Dynamic Gate Fusion' module, Z^{l-1}, $\tilde{Z}^{l}$, Z^{l} denote the feature input, fusion module output, and MLP output, respectively, and $\mathcal{F}_{\text{row}}$, $\mathcal{F}_{\text{col}}$ denote the row-selective scan sequence and column-selective scan sequence, respectively. As shown in Fig. 2(c), the BSS Mamba block consists of two basic elements: a BSS Mamba and an MLP.

3.1.4 Computational Complexity

Assuming that the size of the input feature mapping is $H \times W$, C is the input feature dimension, and N is the state expansion factor of mamba2, the computational complexity of the BSS mamba block is:

$$\Omega\left(\text{BSS} - \text{Mamba}\right) = 2HWN^2 + 12HWC^2 + 8HWC \tag{15}$$

where $12HWC^2$ is the computational complexity of DGF and MLP.

3.2 Overall Architecture

As with the S2WAT architecture, we use UST's traditional encoder-transfer-decoder design. The architecture employs a single encoder to handle the encoding of both content images and style images. An overview is shown in Figure 2.

3.2.1 Encoder

The StyleMamba model employs a patch segmentation module that splits the content and style images into non-overlapping 2×2 patches, each containing 12 values. The linear embedding

layer transforms these patches into a user-defined dimension C. We use CAPE-like learnable positional encoding parameters (18×18 initial size), and the positional encoding is summed with the features by bilinear interpolation adapted to the patch resolution. The embedded patches are processed through BSS Mamba blocks, and these operations constitute the 'first stage'. To realize multi-scale features, StyleMamba introduces downsampling blocks, which extend the number of channels from C to $4C$ by splicing adjacent 2×2 patch sets, and then compressing them to $2C$ by linear projection, while the spatial resolution is reduced to $\frac{1}{2}$ that of the inputs. Subsequently, while maintaining the $\frac{H}{4} \times \frac{W}{4}$ resolution, it passed through the BSS Mamba block, which is 'Stage 2'. This process is repeated in 'Stage 3'. The output resolution is $\frac{H}{8} \times \frac{W}{8}$.

3.2.2 Transfer Module

The transfer module is replaced with a multi-layer transformer de coder, which is identical to the S2WAT [25] model. The precise structure is illustrated in Figure 2(e).

3.2.3 Decoder

To obtain the final result, we adopt the three-layer CNN decoder proposed in [29]. In each layer, we extend the scale by performing a series of operations, including 3×3 Conv + ReLU + 2 × Upsample. Finally, we can save the resultant image in the form of H×W×3.

3.3 Network Optimization

The generated images are subject to original content structure and reference style fusion. Therefore, we construct a content-aware loss $\mathcal{L}_c$ and a style-aware loss $\mathcal{L}_s$ reference [16] respectively, which are used to measure the content image I_c difference and style image I_s difference with the stylized image I_{cs}. The content-aware loss is shown below:

$$\mathcal{L}_c = \sum_{l \in \mathcal{C}} \left\| \phi_l \left(I_{cs} \right) - \phi_l \left(I_c \right) \right\|_2 \tag{16}$$

where ϕ_l denotes the features extracted from the lth layer in the pre-trained VGG19, and the set $\mathcal{C}$ contains the relu4_1 and relu5_1 layers in VGG19. The style-aware loss is defined as:

$$\begin{aligned} \mathcal{L}_s = \sum_{l \in \mathcal{S}} & \left\| \mu\big(\phi_l(I_{cs})\big) - \mu\big(\phi_l(I_s)\big) \right\|_2 \\ & + \left\| \sigma\big(\phi_l(I_{cs})\big) - \sigma\big(\phi_l(I_s)\big) \right\|_2 \end{aligned} \tag{17}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and variance of the features, respectively; the set $\mathcal{S}$ contains the relu2_1, relu3_1, relu4_1 and relu5_1 layers of VGG19.

In our study, we introduced an identity loss function in order to more effectively maintain the structural properties of content images and the stylistic features of stylized images [4]. Two different definitions of identity loss are shown below:

$$\mathcal{L}_{id1} = \left\| I_{cc} - I_c \right\| + \left\| I_{ss} - I_s \right\| \tag{18}$$

$$\mathcal{L}_{id2} = \sum_{l \in \mathcal{N}} \left\| \phi_l \left(I_{cc} \right) - \phi_l \left(I_c \right) \right\|_2 + \left\| \phi_l \left(I_{ss} \right) - \phi_l \left(I_s \right) \right\|_2 \tag{19}$$

where I_{cc} and I_{ss} denote the stylized output images from two identical content images and two identical style images, respectively, and the set N contains the relu2_1 , relu3_1, relu4_1, and relu5_1 layers from VGG19. Finally, the whole network is optimized by minimizing the loss function defined as:

$$\mathcal{L}_{\text{total}} = \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s + \lambda_{id1} \mathcal{L}_{id1} + \lambda_{id2} \mathcal{L}_{id2} \tag{20}$$

to mitigate the effect of magnitude differences, we set λ_c, λ_s, λ_{id1}, and λ_{id2} as the weights of different losses to 2, 3, 50, 1, respectively.

4 Experiments

MS-COCO [18] and WikiArt [1] were used as our content dataset and style dataset, respectively. The batch size was set to 8, and we trained our network with 20,000 iterations. Our model was trained on an NVIDIA 3090 for about 6 hours, and other specific details can be found in our code.

4.1 Style Transfer Results

In this section, we compare our method with Adain[10], SaMam [30], IEST [3], StyTr2 [6], CAP-VSTNet [24], StyleID [4], and S2WAT [25] for comparison.

4.1.1 Qualitative Comparison

In Figure 4, we show the visual results of qualitative comparisons. AdaIN achieves fast style transfer through alignment of mean and variance but has limitations in capturing complex style patterns. IEST generates images with high overall visual quality, yet the generated results may not exactly match the input style references. CAP-VSTNet achieves good performance in maintaining the structure of the content, but local details are lost when complex styles are introduced. StyleID expresses styles well, but content details are not kept intact; furthermore, due to its sensitivity to input size, here we compare the corresponding local detail features. SaMam imposes

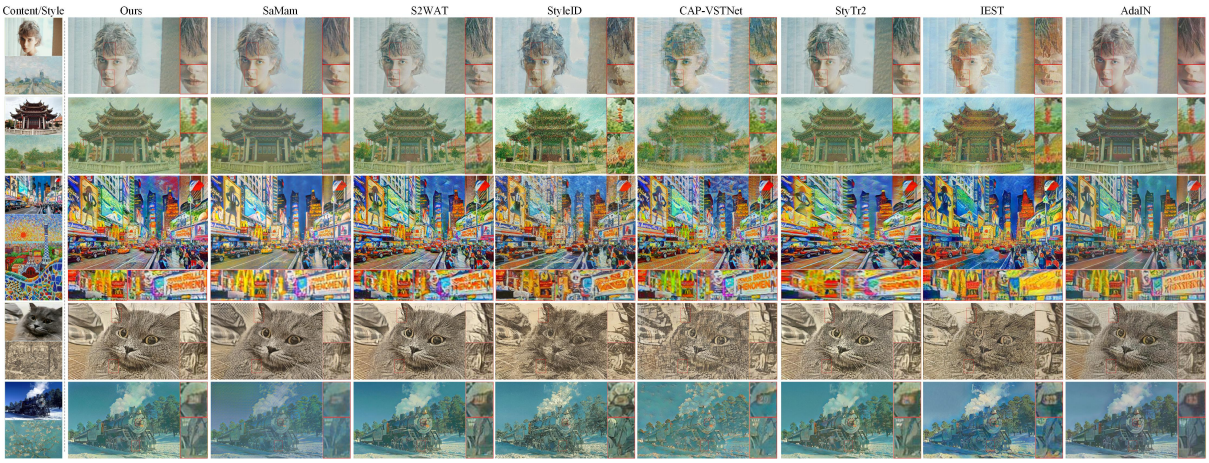


Fig. 4 Qualitative comparisons of style transfer results using different methods.

Table 1 Quantitative evaluation results of different methods. Best results are highlighted in bold, second-best results are underlined.

Method	Our	SaMam	S2WAT	StyleID	CAP-VSTNet	StyTr2	IEST	AdaIN
Content Loss ↓	<u>1.64</u>	8.43	1.68	1.98	**0.89**	1.89	3.53	2.07
Style Loss ↓	**1.60**	15.40	1.71	2.36	<u>4.21</u>	1.69	4.67	5.23
Identity Loss 1 ↓	**0.15**	17.0	<u>0.16</u>	1.28	0	0.30	1.16	11.14
Identity Loss 2 ↓	**1.01**	4.08	<u>1.20</u>	8.72	0	3.85	8.61	29.39

Table 2 Average inference time (in seconds) of different methods at two output resolutions.

Method	Our	SaMam	S2WAT	StyleID	CAP-VSTNet	StyTr2	IEST	AdaIN
512×512	0.153	0.162	0.251	10.63	0.310	0.330	0.040	0.041
256×256	0.031	0.095	0.058	7.339	0.273	0.051	0.020	0.033

high demands on training resources; under identical training settings (i.e., the same number of iterations and batch size) as other methods, the resulting model exhibits suboptimal stylization quality. StyTr2, S2WAT, and our proposed StyleMamba achieve good preservation of content structure and desirable style representations. Notably, StyleMamba excels at preserving local content details (e.g., the woman's lips, temple lanterns, text on billboards, cat's fur, and numerals on a locomotive), as highlighted in the red dashed box regions of Figure 4.

4.1.2 Quantitative Comparison

In this section, we adopt an evaluation protocol similar to that of [1, 5, 11], using style loss, content loss, and identity loss to assess the style quality, content fidelity, and preservation of input information in the generated images, respectively. We also report the average inference time of each method at both output resolutions. As shown in Tables 1, CAP-VSTNet is based on an invertible network and possesses the property of unbiased reconstruction; to reflect this characteristic, we set its identity loss to 0. Despite being unbiased, CAP-VSTNet is outperformed by StyleMamba in terms of both style loss and inference time. StyleMamba achieves the second-lowest content loss and the lowest style loss and identity loss, indicating its strong capability in capturing artistic style features while effectively preserving the structure of the input content.

S2WAT and StyleMamba exhibit comparable loss performance, yet StyleMamba demonstrates a clear advantage in content fidelity. Overall, StyleMamba strikes a favorable balance between preserving the original content structure and incorporating distinctive artistic stylization. Furthermore, as shown in Tables 2, StyleMamba also exhibits superior inference efficiency, outperforming most competing models at both output resolutions. This dual strength—excellent stylization quality combined with high inference speed—establishes StyleMamba as one of the leading approaches in the field of neural style transfer.

4.2 Content Leak

The content leakage problem commonly arises during the style transfer process and is particularly prone to occur when the same stylized image



Fig. 5 Visualization of the content leak: The top/bottom images show the results after running the same stylization process using a certain method after the 1st and 20th rounds, respectively

is repeatedly used for stylization—especially when the model exhibits limited capacity in preserving content details. Following the evaluation protocol in [6, 22], we conduct an in-depth investigation into content leakage during stylization, with a specific focus on StyleMamba, and compare it against S2WAT, StyTr2, StyleID, as well as CNN-based approaches. As shown in Figure 5, our experimental results demonstrate that over 20 iterative stylization steps, StyleMamba and S2WAT incur the least degradation in content details, significantly outperforming both CNN-based methods and SDM-based methods, which often suffer from severe blurring. Although S2WAT alleviates content leakage to some extent, its stylization quality remains inferior to that of StyleMamba. In contrast, StyleMamba not only effectively mitigates content leakage but also achieves high-quality style transfer while faithfully preserving the original content structure and fine details. In summary, StyleMamba offers a novel solution to neural style transfer by simultaneously addressing the content leakage issue and enhancing stylization fidelity.

4.3 Ablation Study

4.3.1 Bidirectional Selective Scan

In order to validate the effectiveness and superiority of "Bidirectional Selective Scan", we conduct a comparative analysis between the row-selective and column-selective scanning strategies within StyleMamba and the original scanning mechanism of Vmamba. The results are shown in Figure 6 (from left to right: content image, style image, row-selective scan, column-selective scan, and bidirectional selective scan). As shown,

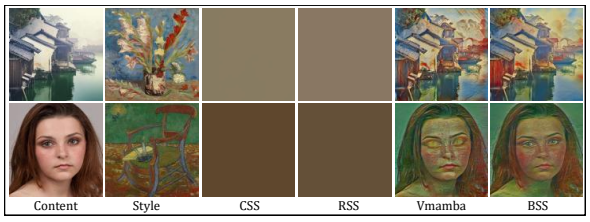


Fig. 6 Visual comparison of using BSS, CSS, Vmamba and RSS

the unidirectional selective scan (either row-only or column-only) fails to generate high-quality stylized images, as it cannot effectively model long-range dependencies. In contrast, our "Bidirectional Selective Scan" successfully establishes connections between local and global contexts, thereby preserving finer content details compared to Vmamba.

4.3.2 Dynamic Gate Fusion

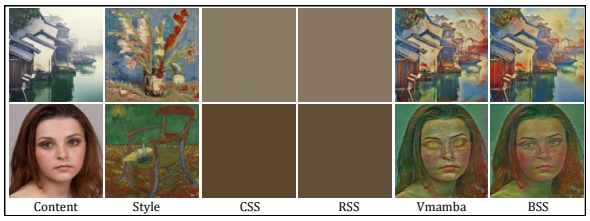


Fig. 7 Visual comparison using 'Sum', 'Concat', and 'DGF'

In order to verify the effectiveness of Dynamic Gate Fusion, we conducted comparison experiments with sum and concat, respectively. As shown in Figure 7 (on the far right, from top to bottom, there are content images, stylized images generated using concat, and stylized images generated using DGF.), the stylized images cannot be generated when using sum. For concat, although a stylized image can be generated, the local details of the generated stylized image have higher fidelity and the stylized effect is more prominent when using our DGF (e.g., areas of snow on mountain tops, window spacing in tall buildings, etc.). The stylized images generated using DGF are satisfactory and produce better-stylized images compared to sum and concat.

4.4 User Study

We further conduct a user study to evaluate the subjective visual quality of the generated results. As shown in Tables 3, we compare our StyleMamba against S2WAT, StyleID, CAP-VSTNet, StyTr2, IEST, and AdaIN. Thirty participants from diverse occupations and age groups are invited to take part in the evaluation. For each test sample, participants are presented with a content image, a style reference image, and the stylized outputs produced by all methods. They are then asked to select their preferred result across three dimensions: (1) overall stylization quality, (2) preservation of content structure, and (3) consistency in transferring stylistic patterns. A total of 300 votes are collected for each dimension. As reported in Table 3.3, StyleMamba achieves the highest vote share (21%) in overall stylization quality, outperforming S2WAT and StyTr2. In terms of content structure preservation,

Table 3 User study results (percentage of votes for each method).

Method	Overall	Content	Style
Our	21%	21%	17%
S2WAT	19%	17%	15%
StyleID	15%	9%	20%
CAP-VSTNet	10%	24%	13%
StyTr2	17%	14%	14%
IEST	11%	8%	13%
AdaIN	7%	7%	8%

StyleMamba obtains 21% of the votes, ranking second among all methods—slightly below CAP-VSTNet's 24%. Regarding consistency of style pattern transfer, StyleMamba demonstrates stable performance with 17% of the votes. These results indicate that StyleMamba strikes a favorable balance between expressive stylization and faithful content retention, exhibiting superior comprehensive performance in subjective evaluation compared to existing approaches.

5 Conclusion

StyleMamba employs a multi-layer Vision Mamba as the encoder backbone and introduces Bidirectional Selective Scan (BSS) within each BSS Mamba block to model spatial dependencies. Specifically, the row-selective scan is more effective at aggregating cross-row contextual information, while the column-selective scan better preserves local structural details and establishes long-range dependencies. These two pathways provide complementary representations, which are adaptively fused through a dynamic gating mechanism to balance global semantics and local textures. This design enables more stable style injection and more faithful preservation of fine details. In terms of efficiency, the scanning computation in BSS exhibits favorable complexity characteristics. Combined with a lightweight fusion module, StyleMamba reduces inference overhead while maintaining high stylization quality, thereby improving the overall efficiency of stylized image generation.

6 Declarations

6.1 Funding

Not applicable.

References

[1] An, J., Huang, S., Song, Y., Dou, D., Liu, W., Luo, J.: ArtFlow: Unbiased image style transfer via reversible neural flows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 862–871 (2021)

[2] Chen, D., Yuan, L., Liao, J., Yu, N., Hua, G.: StyleBank: An explicit representation for neural image style transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1897–1906 (2017)

[3] Chen, H., Wang, Z., Zhang, H., Zuo, Z., Li, A., Xing, W., Lu, D.: Artistic style transfer with internal-external learning and contrastive learning. In: Advances in Neural Information Processing Systems, vol. 34, pp. 26561–26573 (2021)

[4] Chung, J., Hyun, S., Heo, J.P.: Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8795–8805 (2024)

[5] Dao, T., Gu, A.: Transformers are SSMs: Generalized Models and Efficient Algorithms through Structured State Space Duality. Preprint at https://arxiv.org/abs/2405.21060 (2024)

[6] Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: StyTR2: Image style transfer with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11326–11336 (2022)

[7] Dong, W., Zhu, H., Lin, S., Luo, X., Shen, Y., Liu, X., et al.: FusionMamba for Cross-Modality Object Detection. Preprint at https://arxiv.org/abs/2404.09146 (2024)

[8] Efros, A.A., Freeman, W.T.: Image quilting for texture synthesis and transfer. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 341–346 (2001)

[9] Gatys, L.A., Ecker, A.S., Bethge, M.: A Neural Algorithm of Artistic Style. Preprint at https://arxiv.org/abs/1508.06576 (2015)

[10] Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., Re, C.: Combining recurrent, convolutional, and continuous-time models with linear state space layers. In: Advances in Neural Information Processing Systems, vol. 34, pp. 572–585 (2021)

[11] Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1501–1510 (2017)

[12] Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision – ECCV 2016. Lecture Notes in Computer Science, vol. 9906, pp. 694–711. Springer, ??? (2016)

[13] Ju, M., Xie, S., Li, F.: Improving skip connection in U-Net through fusion perspective with Mamba for image dehazing. IEEE Transactions on Consumer Electronics (2024)

[14] Kong, X., Deng, Y., Tang, F., Dong, W., Ma, C., Chen, Y., et al.: Exploring the temporal consistency of arbitrary style transfer: A channelwise perspective. IEEE Transactions on Neural Networks and Learning Systems (2023)

[15] Li, G., Cheng, B., Cheng, L., Xu, C., Sun, X., Ren, P., *et al.*: Arbitrary style transfer with semantic content enhancement. In: Proceedings of the 18th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry, pp. 1–8 (2022)

[16] Li, Z., Pan, H., Zhang, K., Wang, Y., Yu, F.: MambaDFuse: A Mamba-Based Dual-Phase Model for Multi-Modality Image Fusion. Preprint at https://arxiv.org/abs/2404.08406 (2024)

[17] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., *et al.*: VMamba: Visual state space model. In: Advances in Neural Information Processing Systems, vol. 37, pp. 103031–103063 (2024)

[18] Ma, Y., Zhao, C., Li, X., Basu, A.: RAST: Restorable arbitrary style transfer via multi-restoration. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 331–340 (2023)

[19] Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5880–5888 (2019)

[20] Pei, X., Huang, T., Xu, C.: EfficientVMamba: Atrous selective scan for light weight visual Mamba. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 6443–6451 (2025)

[21] Ruan, J., Li, J., Xiang, S.: VM-UNet: Vision Mamba U-Net for Medical Image Segmentation. Preprint at https://arxiv.org/abs/2402.02491 (2024)

[22] Schreiber, S., Geldenhuys, J., De Villiers, H.: Texture synthesis using convolutional neural networks with long-range consistency and spectral constraints. In: 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), pp. 1–6 (2016)

[23] Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance Normalization: The Missing Ingredient for Fast Stylization. Preprint at https://arxiv.org/abs/1607.08022 (2016)

[24] Wen, L., Gao, C., Zou, C.: CAP-VSTNet: Content affinity preserved versatile style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18300–18309 (2023)

[25] Zhang, C., Xu, X., Wang, L., Dai, Z., Yang, J.: S2WAT: Image style transfer via hierarchical vision transformer using strips window attention. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 7024–7032 (2024)

[26] Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10146–10156 (2023)

[27] Zhang, Y., Tang, F., Dong, W., Huang, H., Ma, C., Lee, T.Y., Xu, C.: Domain enhanced arbitrary image style transfer via contrastive learning. In: ACM SIGGRAPH 2022 Conference Proceedings, pp. 1–8 (2022)

[28] Zheng, S., Gao, P., Zhou, P., Qin, J.: PuffNet: Efficient style transfer with pure content and style feature fusion network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8059–8068 (2024)

[29] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., *et al.*: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6881–6890 (2021)

[30] Liu, H., Wang, L., Zhang, Y., *et al.*: SaMam: Style-aware state space model for arbitrary image style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision

and Pattern Recognition (CVPR), pp. 28468–
28478 (2025)