This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

HSI: A Holistic Style Injector for Arbitrary Style Transfer

Shuhao Zhang[1,2], Hui Kang[1,3], Yang Liu[1], Fang Mei[1,3*], Hongjuan Li[1]
[1] College of Computer Science and Technology, Jilin University, Changchun, 130012, China
[2] Academic Affairs Office, Jilin University of Arts, Changchun, 130021, China
[3] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, 130012, China
shzhang22@mails.jlu.edu.cn, kanghui@jlu.edu.cn, liuy24@mails.jlu.edu.cn, meifang@jlu.edu.cn, hongjuan23@mails.jlu.edu.cn

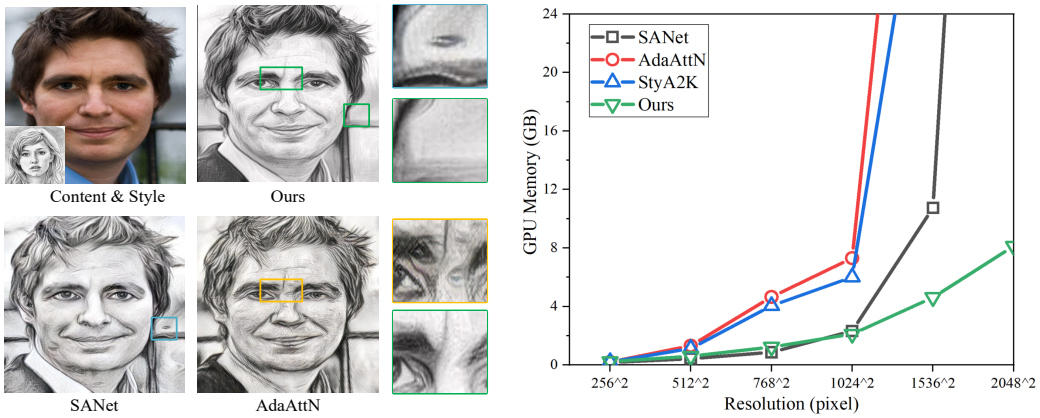


Figure 1. The comparison between our method and some attention-based methods. (a) Image stylization results. Compared to SANet [23] and AdaAttN [21], the generated image of our method is more consistent with original content and more harmonious in style patterns. (b) GPU memory consumption in different resolutions. Compared to other methods, our method successfully renders images from 256 $\times$ 256 to 2048 $\times$ 2048 resolution on a 24GB GPU (4090Ti) without running out of memory.

Abstract

Attention-based arbitrary style transfer methods have gained significant attention recently due to their impressive ability to synthesize style details. However, the point-wise matching within the attention mechanism may overly focus on local patterns such that neglect the remarkable global features of style images. Additionally, when processing large images, the quadratic complexity of the attention mechanism will bring high computational load. To alleviate above problems, we propose Holistic Style Injector (HSI), a novel attention-style transformation module to deliver artistic expression of target style. Specifically, HSI performs stylization only based on global style representation that is more in line with the characteristics of style transfer, to avoid generating local disharmonious patterns in stylized images. Moreover, we propose a dual relation learning mechanism inside the HSI to dynamically render images by leveraging semantic similarity in content and style, ensuring the stylized images preserve the original content and improve style fidelity. Note that the proposed HSI achieves linear computational complexity because it establishes feature mapping through element-wise multiplication rather than matrix multiplication. Qualitative and quantitative results demonstrate that our method outperforms state-of-the-art approaches in both effectiveness and efficiency.

1. Introduction

Arbitrary Style Transfer (AST) is an artwork creating technique to transform a natural image (i.e., content image) with specific artistic semantic by learning the representative style elements of the artistic images (i.e., style image). Since Gates et al. [11] proposed the first neural style transfer algorithm that integrates the Gram statistics of style

*Corresponding author

image into content image, previous methods have demonstrated impressive stylization results by adopting global feature distribution alignment methods [15, 19, 36] and local patch matching strategies [5, 25, 35]. However, these methods can not capture the semantic correspondences between content and style inputs and usually merge style elements with content features directly, thus leading to distorted regions or texture patterns.

To address these problems, researchers have proposed attention-based AST methods [6, 21, 23, 32], to establish the point-to-point semantic correlation between content and style features by equipping the attention module [29]. Due to the adaptive matching of content features and style features, attention mechanism enabled methods thus demonstrating impressive stylization performance, whereas there are still three notable limitations: (1) It is intuitively challenging to deliver representative style patterns only with a single-point feature. The attention mechanism establishes the semantic relations between content and style features at a point-wise level, the holistic style representations are neglected. (2) The exponential calculation of the softmax function may produce biased patterns in the resulting images. Specifically, the attention module may overly concentrate on a prominent style area and ignore the holistic distribution, which may lead to disharmonious patterns. As shown in Figure 1(a), the eye pattern of the style image appears in multiple regions in the stylized image for SANet [23] and AdaAttN [21]. (3) The matrix multiplication will cause quadratic complexity. As shown in Figure 1(b), with the growth of the image resolution, the memory consumption of the attention-based approaches (SANet [23], AdaAttN [21] and StyA2K [38]) increases sharply and ultimately cause GPU memory overflow (24GB on 4090Ti). Therefore, the original attention module for the AST task is incompetent to learn the holistic style representations and is inefficient for end-to-end style transfer based on the above drawbacks.

A new all-to-key attention named StyA2K [38] is proposed to solve these problems recently. It implements a patch-wise aggregation strategy that uses a block region of features instead of a point feature to represent style patterns. While StyA2K puts more style elements to interact with the content feature, it is still challenging in conveying a cohesive style pattern, especially for the artistic image that prioritize style integrity. Moreover, although the patch-wise calculation strategy has reduced the computational load, its complexity still remains at a quadratic level.

To overcome the above challenges, in this paper, we propose a light transfer module called Holistic Style Injector (HSI) to achieve effective and efficient style transfer. The proposed HSI has three traits: (1) Global styles extraction: Instead of calculating point-to-point similarity to guide style transfer, HSI extracts a range of global style statistical features and uses them to establish a straight connection with content features. This operation can render sufficiently comprehensive style characteristics onto the content while avoiding style bias caused by focusing on local areas overly. (2) Dynamic dual relations construction: HSI constructs local-content-to-global-style and global-content-to-global-style relations simultaneously to improve the stylization quality. When the content image is semantically close with the style image, the global-content-to-global-style relation will strengthen, enhancing the overall harmony of the stylized image. (3) Linear-complexity transfer process: HSI employs element-wise multiplication to establish content-style relations, which can perform the transfer process in linear complexity. This multiplication dramatically reduces the computational complexity compared with the matrix multiplication in the attention mechanism. To our knowledge, this is the first study that successfully exploits the power of element-wise multiplication in the style transfer. In summary, the main contributions of this paper are listed as follows:

- This is the first work to propose element-wise multiplication as a fundamental component for building semantic relations between content and style features in arbitrary style transfer. Based on this, we introduce a novel style synthesis module named HSI that is structurally similar to self-attention, yet it produces higher-quality stylized images only with linear computation processes.
- We propose to incorporate multiple types of style statistical features into content features, and establish the dynamical relationship of local-content-to-global-style and global-content-to-global style. These strategies enrich the diversity of stylistic elements in stylized images and improve their style fidelity.
- Extensive experiments demonstrate the excellent performance of our method in effectiveness and efficiency for AST task.

2. Related Work

2.1. Arbitrary Style Transfer

Since Gaty et al. [11] introduced the first neural style transfer model using pre-trained neural networks to synthesize artistic styles through iterative optimization, a variety of arbitrary style transfer algorithms [1, 4, 12, 25, 31, 37] have been developed. Generally, AST methods can be divided into two categories: Global transformation based methods [7, 15, 18, 19, 36] and local matching based methods [5, 6, 14, 23, 25]. For the former, their common idea is to directly match the style statistical distribution at the global feature level. Typical examples are AdaIN [15] and



Figure 2. The network framework of our method.

WCT [19]. Specifically, AdaIN directly matches the mean and variance between content and style features at the channel level. Similarly, WCT introduces whitening and coloring operations to align the overall style distribution. While these methods are efficient for stylization, they often struggle with visual quality because they can not capture the semantic relationship content and style elements.

For the latter, they emphasize the local semantic alignment between content features and style features. StyleSwap [5] and Avatar-Net [25] are early representatives of this class of methods. They fuse style elements by calculating the similarity between content features and style features on a patch-wise level, often achieving superior visual quality compared to the global transformation based methods. With self-attention [29] gaining popularity across various fields, many researchers began exploring attention-based AST methods [4, 21, 23]. These methods can render the content image according to the point-wise semantic similarity learned adaptively between content features and style features. SANet [23] directly uses the self-attention module to perform style transfer in the feature space. Based on SANet, AdaAttN [21] further integrates the advantages of AdaIN algorithm for feature normalization. Furthermore, StyTr^2 [6] and S2WAT [32] employ transformer components [10] to extract style features and transfer styles. While attention-based methods can effectively preserve semantic information of content and style, they often overly focus on aggregating local patterns. This may lead to content distortion of the generated images due to the repetitive and scattered style patterns. In addition, the quadratic complexity is a common challenge for the attention-based methods. Recently, StyA2K [38] tries to address above problems by proposing an all-to-key attention module to improve style transfer stability and efficiency. Compared with the point-wise similarity calculation in self-attention, all-to-key attention builds similarity map in a patch-wise feature level. However, the generated results are often under-stylized due to the excessive emphasis on structural stability. Moreover, its computational complexity remains quadratic with the image size despite the improvement in efficiency. In this paper, we aim to explore a more effective transformation module that achieves high-quality stylization with sufficiently low computational complexity.

2.2. Attention Mechanism

The attention mechanism was first introduced in machine translation by Bahdanau et al [2]. The most well-known variant is self-attention, which was introduced in the Transformer model [29]. Self-attention is widely used in natural language processing [8, 9, 27, 28] and computer vision [3, 16, 22, 33] due to its ability to establish long-distance dependencies. Its robust scalability and adaptability in establishing similarity relationships between local features have attracted many researchers to apply it for improving performance in their respective fields. However, in image generation tasks like AST [21, 23], the original dense relational learning mechanism in self-attention may negatively affect visual quality due to focusing on local style areas. Moreover, the high computational complexity of self-attention is also a common challenge. In this paper, we seriously consider these issues and explore how style patterns can better match the content during the feature interaction, and propose an effective and efficient style transformation module for AST task.

3. Proposed Method

3.1. Overall Framework

The overall style transfer network is shown in Figure 2. A simple encoder-decoder architecture is adopted in this paper, where encoder $E(\cdot)$ is a parameter-fixed pre-trained VGG-19 [26] network (up to relu_4_1 layer). After extracting the content feature F_c and style feature F_s, a chain of connected HSI modules will receive both features to complete style transform in feature space, which generate the synthetic feature F_{cs}:

$$F_{cs} = \text{HSIs}(E(I_c), E(I_s)) \tag{1}$$

Then, the decoder will invert the F_{cs} to the image space that produces the stylized image I_{cs}:

$$I_{cs} = D(F_{cs}) \tag{2}$$

where the structure of decoder is an mirror of the encoder. Following [19], to alleviate the checkerboard effects, we adopt the nearest up-sampling to replace the pooling layer. Moreover, the reflection padding is used in encoder and decoder to prevent the board collapse.

3.2. Revisit Attention Mechanism in AST

In AST, attention mechanism refers to the self-attention [29] by default. Benefiting from flexible relation building ability, attention module can adaptively embed style patterns into content features of each position for stylization. Figure 3(a) presents the overall encoding process. Query (Q), key (K) and value (V) are the core encoding feature of

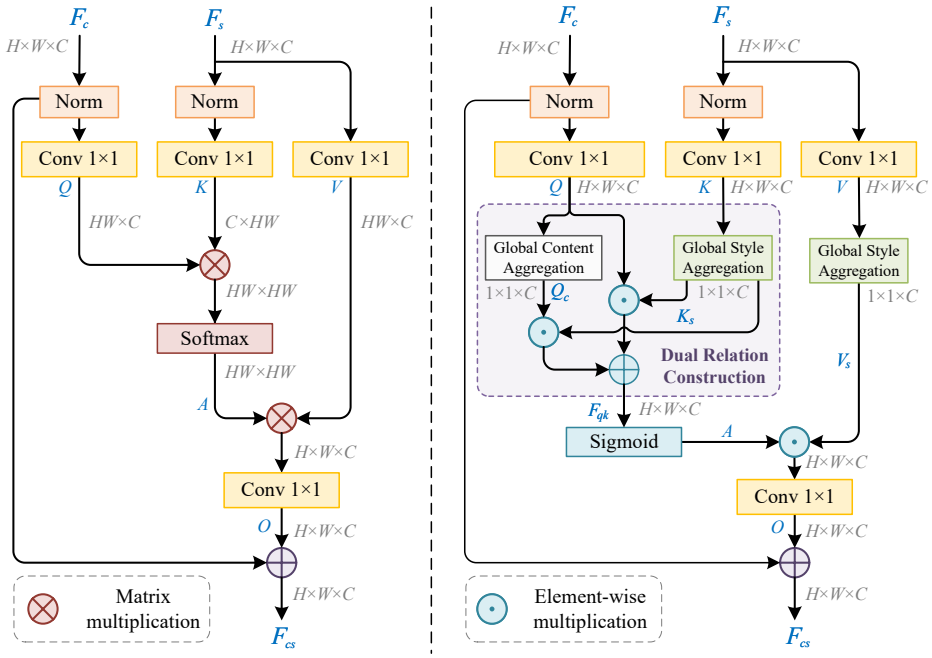


Figure 3. The structure and feature encoding process comparison with self-attention (a) and our HSI module (b). HSI has a similar structure to self-attention, which uses element-wise multiplication instead of matrix multiplication to model the semantic similarity of content features and style features.

attention module, they can be obtained by:

$$Q = f_q(Norm(F_c)), K = f_k(Norm(F_s)), V = f_v(F_s) \tag{3}$$

where f_q, f_k and f_v denote the convolution layer which kernel usually with size of 1×1. $Norm(\cdot)$ is the mean-variance channel-wise normalization. We can obtain the attention score A by

$$A = softmax(Q \otimes K^T) \tag{4}$$

where $\otimes$ is the matrix multiplication. Attention map A indicates the point-to-point similarity score between content feature F_c and style feature F_s. By the similarity map A, the extracted style elements can be transferred on the content features through

$$O = f_o(A \otimes V) \tag{5}$$

where f_o denotes the convolution layer like f_q, f_k and f_v. O is the synthetic feature for generating stylized image, which is also added to the input content feature F_c to improve the structure consistency, as shown in Figure 3(a).

As analyzed in the Introduction, although the attention mechanism establishes a point-to-point matching relationship between content features and style features to improve semantic consistency in stylization results, it may overfocus on a particular style region, leading to an inharmonious pattern in the resulting image. In addition, the attention mechanism has quadratic computational complexity with the size of the input image. The detailed complexity analysis is as follows.

Complexity Analysis. We assume Q, K and V with size of (H, W, C). The computation of self-attention mainly focuses on two phases. One is the dot product in each position between Q and K for generating attention map A that with size of $(H \times W, H \times W)$. Another is the weighted summation of the A and V in Equation 5. The computational complexity of both are $\mathcal{O}((H \times W)^2 \times C)$, so the computational complexity of self-attention is $\mathcal{O}((H \times W)^2 \times C)$.

3.3. Holistic Style Injector

Although the attention mechanism can produce promising stylization results in AST, it also has obvious limitations in terms of style rendering and computational efficiency. To overcome these issues, we propose a transformation module called Holistic Style Injector (HSI), which builds more comprehensive correlations between style and content features for improved stylization while reducing computational complexity. On the whole, HSI and the attention module have similar processing processes, as shown in Figure 3. First, the content feature F_c and the style feature F_s are normalized and linearly processed to obtain the key triplet (Q, K, V) for stylization. Next, Q will interact with K and their results further be normalized probabilistically to generate the attention score A that hides the semantic relationships between the content and style features. Then,

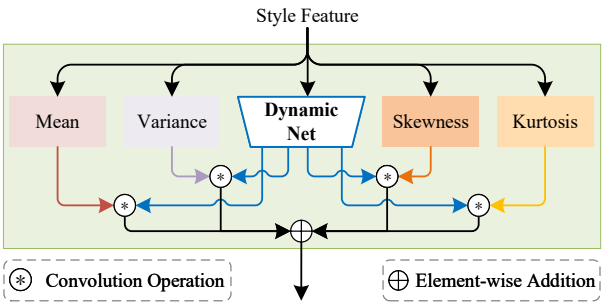


Figure 4. The detailed illustration of global style aggregation in solid green box of Figure 3(b).

the style representation V is weighted and summed with A, followed by a linear transform to synthesize style elements and produce the output feature O. Finally, a residual connection is established to link the input content feature F_c with O, ensuring a more consistent content structure in the stylization results.

There are noticeable differences between HSI and attention module in constructing relationships between content and style features. In contrast with attention module, HSI introduces three key innovations: (1) Global styles extraction, (2) Dynamic dual relations construction, and (3) Linear-complexity transfer process.

Global Styles Extraction. Unlike the attention module that calculates the point-wise feature similarity, HSI focuses on mining the relationship between global style features and content features. To describe the style pattern more comprehensively, we utilize four channel-wise statistical features, including mean, variance, skewness and kurtosis, combining the content features perform the stylization, as shown in Figure 4. Specifically, the mean and variance express the overall tone and variation degree of style elements, while skewness and kurtosis indicate the symmetry and concentration of the distribution of style features. Their definitions are as follows:

$$\begin{aligned}
\mu &= \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} x_{h,w,c} \\
\sigma^2 &= \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} (x_{h,w,c} - \mu)^2 \\
\gamma_1 &= \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(\frac{x_{h,w,c} - \mu}{\sigma} \right)^3 \\
\gamma_2 &= \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(\frac{x_{h,w,c} - \mu}{\sigma} \right)^4
\end{aligned} \tag{6}$$

where μ, σ, γ_1 and γ_2 denote the channel-wise mean, standard deviation, skewness and kurtosis respectively. Then, we employ a dynamic network [17] that transforms average and maximum value of style component K into the weight W and bias b:

$$\begin{aligned}
W &= AvgPool(Conv(K)) \\
b &= MaxPool(Conv(K))
\end{aligned} \tag{7}$$

where $Conv(\cdot)$ denotes the depth-wise separable convolution, $AvgPool(\cdot)$ and $MaxPool(\cdot)$ are global average/max pooling operation. Note that each statistical feature has its own W and b. Then, four style statistical features are combined in a weighted summation manner and the result will interact with the content feature.

Dynamic Dual Relations Construction. To better exploit the style patterns of the original image, we establish two relationships: local-content-to-global-style and global-content-to-global-style, as illustrated in the purple dashed box of Figure 3(b). Specifically, the global style feature K_s interacts with Q, which contains refined content information, and the global content feature Q_c, which encompasses overall information from the Q. This interaction process is achieved through:

$$F_{qk} = \lambda_g \times (Q_c \odot K_s) \oplus (1 - \lambda_g) \times (Q \odot K_s) \tag{8}$$

where $\odot$ and $\oplus$ denote the element-wise multiplication and element-wise addition with broadcast mechanism. The fused feature F_{qk} is obtained based on the similarity coefficient λ_g:

$$\lambda_g = \frac{\frac{Q \cdot K}{\|Q\| \|K\|} + 1}{2} \tag{9}$$

In our idea, we pay more attention to the global semantic relation between content and style feature. Therefore, when both the content image and style image are semantically similar, such as both being human faces, the value of λ_g will be larger. This emphasizes the importance of the global-content-to-global-style relation for stylization, which can help reduce the disharmonious patterns caused by excessively focusing on local style patterns.

Linear-Complexity Transfer Process. The linear complexity of HSI is mainly reflected on the element-wise multiplication operation, which appears in several nodes of the HSI module, as shown in Figure 3(b). When the size of input feature are (H, W, C), the complexity of element-wise multiplication is $\mathcal{O}(H \times W \times C)$ because only the element multiplication at corresponding positions is involved. Therefore, the complexity of HSI module is $\mathcal{O}(H \times W \times C)$.

3.4. Loss Functions

We optimize the entire network using three term of loss functions. The first term is the style loss $\mathcal{L}_s$ as defined in [15], which is responsible for aligning the mean μ and standard deviation σ between the stylized image I_{cs} and the



Figure 5. Qualitative comparisons with state-of-the-art AST methods. Zoom in for a better view.

style image I_s in VGG feature space:

$$\mathcal{L}_s = \sum_{i=1}^{N_l} \|\mu(\phi_i(I_{cs})) - \mu(\phi_i(I_s))\|_2 + \|\sigma(\phi_i(I_{cs})) - \sigma(\phi_i(I_s))\|_2 \tag{10}$$

where ϕ_i denotes lth layer of the pre-trained VGG network and N_l represents the number of layers. The second term is the content loss $\mathcal{L}_c$ for preserving content structure:

$$\mathcal{L}_c = \sum_{i=1}^{N_l} \|\phi_i(I_{cs}) - \phi_i(I_c)\|_2 + \|\phi_i(I_{cs}^{gray}) - \phi_i(I_c^{gray})\|_2 \tag{11}$$

where I_*^{gray} represents the corresponding gray image. In addition, we adopt a adversarial loss $\mathcal{L}_{adv}$ based on a multi-scale discriminator to improve holistic stylization effect.

The three losses are summed to consist the full loss:

$$\mathcal{L} = \lambda_s \mathcal{L}_s + \lambda_c \mathcal{L}_c + \lambda_{adv} \mathcal{L}_{adv} \tag{12}$$

where λ_s, λ_c and λ_{adv} are the weighting factors. We set λ_s, λ_c and λ_{adv} as 60, 5 and 50 to strike a balance among losses.

4. Experiments

More implement details and experimental results can be found in Supplementary Material.

4.1. Implement Details

We train our model using the MS-COCO dataset [20], which contains approximately 120k real photos, and the WikiArt dataset [24], which includes around 80k artistic images. During the training phase, we first resize each training image to 512 × 512 and then randomly crop it to 256 × 256 as the input. We utilize Adam as the optimizer. The batch size and learning rate are set as 4 and 0.0001 respectively. We train our model for 200k iterations on an NVIDIA GTX 4090Ti for about eight hours.

4.2. Comparison with State-of-the-Art Methods

To evaluate the performance of our method, we compare it with six state-of-the-art AST methods: AdaIN [15], SANet [23], AdaAttN [21], StyTr^2 [6], StyA2K [38] and AesPA-Net [14]. Among these, SANet and AdaAttN are the pure attention based methods. We also employ some improved attention based approaches like the StyA2K and AesPA-Net to compare stylization performance. All results of these methods are obtained by running their public released codes with default settings.

Qualitative Evaluation. The comparison in Figure 5 demonstrates the superiority of our method in visual quality. Our stylized results faithfully reflect the style elements such as colors and strokes from the style image, and no local collapse in the stylization results. Additionally, our results maintain an excellent consistency in content structure compared to other methods. In contrast, the generated images of AdaIN [15] lack sharp details and fine brushstrokes because its transfer process does not combine style semantic information. Since the style transformation of SANet [23] and AdaAttN [21] is entirely based on the attention mechanism, they may repeat local salient patterns of style image, such as the repetitive eyes of human faces in the 1st row and the clouds in the 2nd row. StyTr^2 [6] and StyA2K [38] tend to preserve the content structure excessively, which easily results in an under-stylization, such as the face and collar of the man in the 1st row still remain some colors of content image. AesPA-Net [14] can reduce the pattern repetition caused by the attention module, but it also inadvertently erases the brushstrokes that increase the beauty of the

Table 1. Quantitative comparisons with state-of-the-art AST methods. We employ average content loss, style loss, LPIPS and FID metrics to measure the visual quality of the generated images.

Method	**AdaIN** [15]	**SANet** [23]	**AdaAttN** [21]	**StyTr^2** [6]	**StyA2K** [38]	**AesPA-Net** [14]	**Ours**
Content Loss ↓	0.97	1.18	1.21	0.69	**0.59**	0.63	0.62
Style Loss ↓	1.44	1.26	1.52	1.34	1.21	0.99	**0.95**
LPIPS ↓	0.65	0.63	0.57	0.56	0.49	0.52	**0.46**
FID ↓	19.68	18.74	19.34	18.97	19.85	20.07	**18.46**

Table 2. The average inference time (T) and GPU memory consumption (M) on four image resolutions. "OOM" means "cuda out of memory".

Resolution	**AdaIN** [15]		**SANet** [23]		**AdaAttN** [21]		**StyTr^2** [6]		**StyA2K** [38]		**AesPA-Net** [14]		**Ours**	
	T(s)	**M(GB)**	**T(s)**	**M(GB)**	**T(s)**	**M(GB)**	**T(s)**	**M(GB)**	**T(s)**	**M(GB)**	**T(s)**	**M(GB)**	**T(s)**	**M(GB)**
256×256	0.0029	0.13	0.0038	0.17	0.0038	0.21	0.0255	0.41	0.0036	0.21	0.0057	0.67	0.0076	0.23
512×512	0.0030	0.45	0.0048	0.48	0.0487	1.32	0.1326	1.67	0.0038	1.13	0.0503	2.40	0.0081	0.61
1024×1024	0.0035	1.71	0.0051	2.48	0.1538	7.31	OOM	OOM	0.0053	6.01	0.2235	10.06	0.0203	2.11
2048×2048	0.0092	6.75	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	0.1535	8.12

stylized images. In general, our method achieves a better balance in content preservation, stylization depth and visual quality compared to the comparison methods.

Quantitative Evaluation. We use content loss, style loss, LPIPS [34] and FID [13] to evaluate the visual quality of stylized images. Content/style loss measures the content/style consistency between the content/style image and the stylized images. LPIPS and FID evaluate the content/style fidelity between the stylized image and the corresponding content/style image. For these metrics, a lower score indicates better quality of the generated stylizations. We randomly selected 20 content and 20 style images to generate 400 stylized images for comparison. Table 1 shows the experimental results. As we can observe, our method achieves the lowest style loss, LPIPS and FID, and is only slightly higher than StyA2K in content loss, which shows that our method is significantly competitive in both transferring style information and preserving content structure compared with the state-of-the-art methods.

Efficiency Analysis. In Table 2, we compare the inference time and GPU memory consumption at four image resolutions: 256×256, 512×512, 1024×1024 (1K) and 2048×2048 (2K). All experiments are run on an NVIDIA 4090Ti GPU. We can observe that the models containing the attention module all face memory overflow problems when processing 2K resolution. The most obvious is StyTr^2 that based on Transformer architecture, which start overflows from 1K resolution. This is because the computational complexity of these models is quadratic with the size of the input image. In contrast, our HSI module achieves linear complexity with the image size. The results show that our method can run successfully at any test resolution. When processing 1K images, our method achieves about 50 fps and occupies only about 2GB of GPU memory. The results show that our method is capable of performing style transfer in real time.

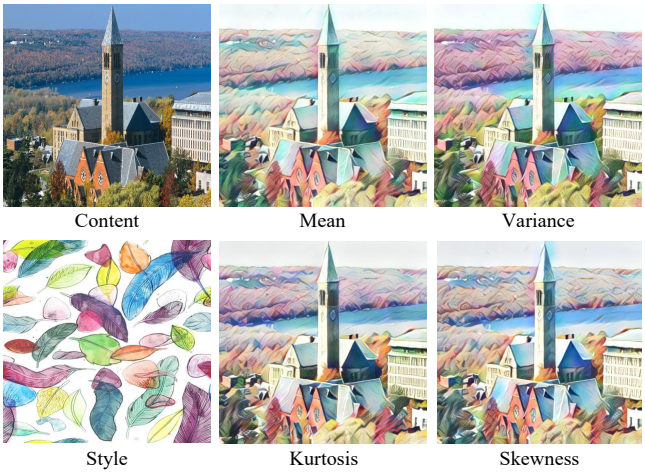


Figure 6. Effectiveness of four style statistic features.

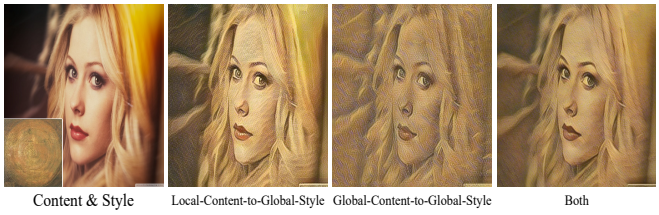


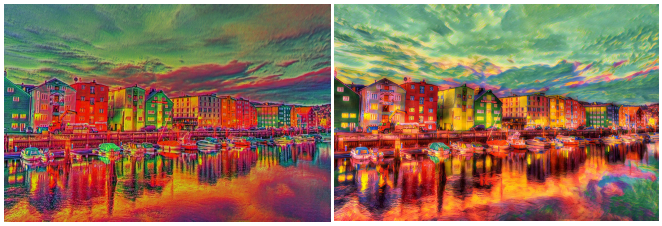
Figure 7. Effectiveness of dual relations construction.

4.3. Ablation Study

Effectiveness of Four Style Statistic Features. In order to verify the effectiveness of four style statistical features (mean, variance, kurtosis and skewness), we interact each style features with the content features separately. As shown in Figure 6, each statistical feature can bring different colors and stroke in different content areas. This ensures that HSI can inject richer style elements into the stylized im-



Content & Style AdaIN



MicroAST Ours

Figure 8. Comparison on high-resolution (2K) style transfer.



Content Style SANet SANet + HSI

Figure 9. Generalization of HSI module. By replacing the self-attention in SANet with our HSI module, the distorted regions and disharmonious patterns in the results of original SANet are reduced significantly, and the semantic structure of content image is well preserved.

age and has more possibilities to improve the visual effect of the resulting image.

Effectiveness of Dual Relations Construction. To verify the effectiveness of the dual relations construction mechanism, we construct only local-content-to-global-style or global-content-to-global-style relation and both in HSI module. As shown in Figure 7, we can observe that the stylized image preserves more content details when only constructing a local-content-to-global-style relation, while constructing global-content-to-global-style relation can better integrate the global style. The resulting image brought by dynamically combining both relations can fuse their respective advantages. This proves that constructing two relationships at the same time can improve the consistency of content and style of the resulting image.

4.4. More Discussions

High-Resolution Image Style Transfer. To verify the stylization effectiveness of our method on high-resolution images, we conducted experiments using 2K resolution images. We compared our proposed method with AdaIN and MicroAST [30]. MicroAST is a professional high-resolution style transfer model. The experiment results are shown in Figure 8. Our method renders style elements such as brushes and colors more abundantly and compared with AdaIN and MicroAST, which shows its effectiveness in transferring holistic style patterns.

Generalization of HSI module. The HSI module proposed in this paper is a plug-and-play module. To verify its generalization performance, we replace the attention module in SANet with the HSI module, denoted as SANet + HSI. We retrain the model with the default settings and compare it with the original SANet. The results are shown in Figure 9. In the results of the original SANet, we can observe obvious overly repetitive style patterns, such as the faces in the 1st row and the red palm prints in the 2nd row, which cause local distortions that greatly reduce the visual quality. In contrast, when the HSI module replaces the attention module, the repeated patterns and distorted areas in the stylized image are significantly reduced. This further demonstrates the effectiveness of the HSI module in both capturing the overall style pattern and preserving the content structure.

5. Conclusion

In this paper, we propose a new attention-style transfer module, named Holistic Style Injector (HSI), for efficient and effective arbitrary style transfer. It contains three characteristics: Global styles extraction, dynamic dual relations construction, and the linear transfer process. Global style extraction emphasizes rendering various global style patterns on the content image while preventing local distortions and enriching style information. Dynamic dual relation construction establishes local-content-to-global-style and global-content-to-global-style relations simultaneously. It flexibly adjusts the matching between style and content distribution according to semantic similarity to obtain more harmonious stylization results. The linear transfer process refers to the style integration in our model based on efficient element-wise multiplication, which can complete stylization in a linear complexity. Extensive experiments demonstrate the superiority of our method.

Acknowledgement

This work is supported in part by the National Key Research and Development Program of China (2022YFB4500600), in part by the Science and Technology Development Plan Project of Jilin Province (20240302079GX).

References

[1] 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. 2

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015. 3

[3] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 3

[4] 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. 2, 3

[5] Tian Qi Chen and Mark Schmidt. Fast patch-based style transfer of arbitrary style. *Workshop in Constructive Machine Learning*, 2016. 2, 3

[6] 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. 2, 3, 6, 7

[7] 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. 2

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186, 2019. 3

[9] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32, 2019. 3

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[11] 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. 1, 2

[12] Shuyang Gu, Congliang Chen, Jing Liao, and Lu Yuan. Arbitrary style transfer with deep feature reshuffle. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8222–8231, 2018. 2

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[14] 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. 2, 6, 7

[15] 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. 2, 5, 6, 7

[16] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023. 3

[17] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016. 5

[18] Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. Dynamic instance normalization for arbitrary style transfer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4369–4376, 2020. 2

[19] 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. 2, 3

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[21] 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. 1, 2, 3, 6, 7

[22] Yaoyao Liu, Bernt Schiele, Andrea Vedaldi, and Christian Rupprecht. Continual detection transformer for incremental object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23799–23808, 2023. 3

[23] 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. 1, 2, 3, 6, 7

[24] Fred Phillips and Brandy Mackintosh. Wiki art gallery, inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3):593–608, 2011. 6

[25] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 8242–8250, 2018. 2, 3

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[27] David So, Wojciech Mańke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V Le. Searching for efficient transformers for language modeling. *Advances in neural information processing systems*, 34:6010–6022, 2021. 3

[28] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020. 3

[29] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2, 3

[30] Zhizhong Wang, Lei Zhao, Zhiwen Zuo, Ailin Li, Haibo Chen, Wei Xing, and Dongming Lu. Microast: towards super-fast ultra-resolution arbitrary style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2742–2750, 2023. 8

[31] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14618–14627, 2021. 2

[32] Chiyu Zhang, Xiaogang Xu, Lei Wang, Zaiyan Dai, and Jun Yang. S2wat: Image style transfer via hierarchical vision transformer using strips window attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7024–7032, 2024. 2, 3

[33] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2998–3008, 2021. 3

[34] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[35] Yulun Zhang, Chen Fang, Yilin Wang, Zhaowen Wang, Zhe Lin, Yun Fu, and Jimei Yang. Multimodal style transfer via graph cuts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5943–5951, 2019. 2

[36] 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. 2

[37] 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. 2

[38] Mingrui Zhu, Xiao He, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. All-to-key attention for arbitrary style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23109–23119, 2023. 2, 3, 6, 7