

TSSAT: Two-Stage Statistics-Aware Transformation for Artistic Style Transfer

Haibo Chen*

Nanjing University of Science and Technology
Nanjing, China
hbchen@njust.edu.cn

Jun Li

Nanjing University of Science and Technology
Nanjing, China
junli@njust.edu.cn

Lei Zhao

Zhejiang University
Hangzhou, China
cszhl@zju.edu.cn

Jian Yang

Nanjing University of Science and Technology
Nanjing, China
csjyang@njust.edu.cn

ABSTRACT

Artistic style transfer aims to create new artistic images by rendering a given photograph with the target artistic style. Existing methods learn styles simply based on global statistics or local patches, lacking careful consideration of the drawing process in practice. Consequently, the stylization results either fail to capture abundant and diversified local style patterns, or contain undesired semantic information of the style image and deviate from the global style distribution. To address this issue, we imitate the drawing process of humans and propose a Two-Stage Statistics-Aware Transformation (TSSAT) module, which first builds the global style foundation by aligning the global statistics of content and style features and then further enriches local style details by swapping the local statistics (instead of local features) in a patch-wise manner, significantly improving the stylization effects. Moreover, to further enhance both content and style representations, we introduce two novel losses: an attention-based content loss and a patch-based style loss, where the former enables better content preservation by enforcing the semantic relation in the content image to be retained during stylization, and the latter focuses on increasing the local style similarity between the style and stylized images. Extensive qualitative and quantitative experiments verify the effectiveness of our method.

CCS CONCEPTS

- Applied computing → Fine arts; • Computing methodologies → Appearance and texture representations; *Image manipulation*.

KEYWORDS

artistic style transfer, global statistics alignment, local statistics swap, attention

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-0108-5/23/10...\$15.00
<https://doi.org/10.1145/3581783.3611819>

ACM Reference Format:

Haibo Chen, Lei Zhao, Jun Li, and Jian Yang. 2023. TSSAT: Two-Stage Statistics-Aware Transformation for Artistic Style Transfer. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3581783.3611819>

1 INTRODUCTION

Artistic style transfer is a powerful technique for image editing and art creation, whose key problem is how to separate and recombine the contents and styles of given images. Recently, the seminal work of Gatys *et al.* [12] firstly proposed to leverage a pre-trained Deep Convolutional Neural Network (DCNN) to tackle this problem, which opens up the neural style transfer era. Since then, numerous neural style transfer methods have been developed. Among them, global statistics-based and local patch-based methods dominate the current style transfer field.

More specifically, global statistics-based methods [12, 14, 15, 17, 18, 31, 42, 55] focus on exploring proper global statistics to represent style and enforcing the global statistics of the content image to be aligned with those of the style image for stylization. For example, Huang *et al.* [15] found that the mean and variance of deep image features carried style information, and Li *et al.* [30] employed the whitening and coloring transforms to reflect the direct matching of feature covariance of the content image to a given style image. This line of work is able to align the global style distributions between the style and stylized images. However, they overlook one critical problem: a style image usually contains more than one kind of style pattern, and similarly, a content image always consists of many different semantic regions. Simply transferring the global style to the content image considers neither the diversity of local style patterns nor the difference among multiple content regions (e.g., AdaIN [15] fails to capture the abundant style information in the 1st row of Figure 1). In contrast, local patch-based methods [3, 7, 27, 35, 38, 44, 51, 53] conduct style transfer by replacing every content patch with similar style patches in the feature space. For example, Chen *et al.* [7] performed a style-swap operation that swaps each content feature patch with its closest-matching style feature patch, and Park *et al.* [35] introduced a style-attentional network that integrates the style feature patches according to the semantic spatial distribution of the content image. Despite the effectiveness in learning local style patterns, this line of work usually

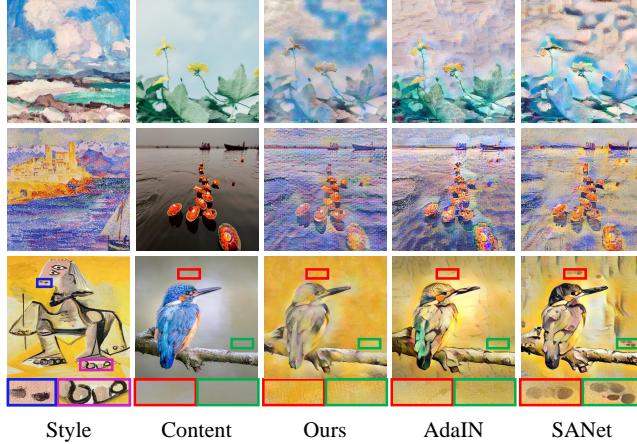


Figure 1: Stylization examples. The first two columns show the style and content images. The other three columns show the stylized images produced by our method, AdaIN [15], and SANet [35].

suffers from two drawbacks: 1) the style feature patches will inevitably introduce some semantic information of the style image to the stylization result (e.g., the stylized image generated by SANet [35] in the 3rd row of Figure 1 contains the structure of the nose and toes of the style image); 2) some marginal colors and texture patterns in the style image may prevail in the stylization result, while those critical colors and texture patterns in the style image may be ignored. This is because these methods only consider the semantic correspondence between the content and style images, neglecting the global style distribution of the style image (e.g., the 2nd row, 5th column of Figure 1).

Motivated by the observations and analyses above, we propose a Two-Stage Statistics-Aware Transformation (TSSAT) module, which **simulates the drawing process of humans**, i.e., first drawing the basic and primary structures and textures from a global perspective and then further enhancing the paintings with delicate fine-grained details from a local perspective. Our TSSAT accomplishes a similar process with a global statistics alignment stage and a local statistics swap stage. In detail, TSSAT first aligns the global statistics of content and style features to learn global style, inspired by global statistics-based methods [15, 17, 19, 34, 48]. Then, for each local patch of the content feature, TSSAT finds its closest-matching style feature patch and **swaps their local statistics instead of the two local patches**, which overcomes the inertial thinking of local patch-based methods [7, 28, 35, 46, 51] and revolutionizes local style learning. In this way, our method prevents the semantic information stored in local style patches from being introduced into the stylization result. Meanwhile, more abundant and fine-grained local style patterns are involved on the basis of learned global style distribution. Our TSSAT also allows flexible style pattern modulation by adjusting the patch size in the local statistics swap stage. Moreover, to further enhance both content and style representations, we introduce two novel losses: an attention-based content loss and a patch-based style loss. To be more specific, the attention-based content loss enforces the semantic relation in the

content image to be retained during stylization, leading to better content preservation. And the patch-based style loss focuses on increasing the style similarity between the style and stylized images from a local perspective.

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

- We propose a TSSAT module that harnesses feature statistics to first build the global style foundation and then enrich local style details, significantly improving the stylization effects and providing fresh insight into the challenging style transfer problem.
- An attention-based content loss is introduced to enable better content preservation by enforcing the semantic relation in the content image to be retained during stylization.
- A patch-based style loss is introduced to increase the style similarity between the style and stylized images from a local perspective.
- Comprehensive experimental results show that our method outperforms state-of-the-art style transfer methods both qualitatively and quantitatively.

2 RELATED WORK

Global Statistics-based Style Transfer. Global statistics-based methods generally transform the content features to match the global statistics of style features for stylization. Gatys *et al.* [12] represented the style of an image with Gram matrix and constrained the Gram matrices of the style and stylized images to be consistent by iterative optimizations. Huang *et al.* [15] performed style transfer by adjusting the mean and variance of the content features to match those of the style features. Li *et al.* [30] conducted the whitening and coloring transforms (WCT) to endow the content features with the same statistical characteristics as the style features. Instead of directly using the first- or second-order statistical transformation to learn style, Li *et al.* [29] employed light-weighted CNNs to predict a learnable linear transformation matrix conditioned on an arbitrary pair of content and style images. Jing *et al.* [17] extended the work of Huang *et al.* [15] by introducing a dynamic instance normalization (DIN) module that encodes a style image into learnable convolution parameters, upon which the content image is stylized. An *et al.* [1] presented an unbiased style transfer framework that consists of reversible neural flows [13] and an unbiased style transfer module (e.g., AdaIN [15] or WCT [30]) to address the content leak problem. Lin *et al.* [31] proposed a drafting network and a revision network to perform style transfer in a progressive procedure and relied on AdaIN [15] to combine the style feature and the content feature. Recently, some methods [5, 6, 24, 25, 37, 49, 56] proposed to learn style from a collection of artworks rather than a single style image, vastly improving the quality of stylization results. The above methods significantly promote the development of style transfer. However, given that the style image usually contains more than one kind of style patterns and the content image always consists of multiple different semantic regions, it may be insufficient to use such global statistics to represent style.

Local Patch-based Style Transfer. Local patch-based methods generally swap local content patches with similar local style patches in the feature space for stylization. Chen *et al.* [7] concatenated both

content and style information into a single layer of the CNN, by swapping the textures of the content image with those of the style image. Sheng *et al.* [38] proposed a patch-based feature manipulation module to transfer the content features to semantically nearest style features. Park *et al.* [35] and Deng *et al.* [10] embedded a local style pattern in each position of the content features by mapping a relationship between the content and style features based on attention mechanism. Zhang *et al.* [53] clustered the style image features into sub-style components, which are matched with local content features under a graph cut formulation. Yao *et al.* [51] employed self-attention as a residual to obtain the attention map, and then introduced multi-scale style swap and a stroke fusion strategy to adaptively integrate multiple style patterns into one stylized image. Huo *et al.* [16] proposed a manifold alignment-based style transfer framework which allows semantically similar regions between the output and style images share similar style patterns. Chen *et al.* [3] introduced an internal-external learning scheme and two contrastive losses to bridge the gap between human-created and AI-created artworks. Deng *et al.* [9] proposed a transformer-based [40] style transfer framework that translates the content sequences based on the reference style sequences, leading to stylization results with well-preserved structures. Although these methods are effective in learning more local style patterns, the generated stylized images usually contain undesired semantic information of the style image and sometimes deviate from global style distribution.

Others. Beyond that, a number of methods have been proposed to address the style transfer problem from other perspectives. Liu *et al.* [33] integrated the ideas of Huang *et al.* [15] and Park *et al.* [35] and proposed an attention and normalization module, named AdaAttN, which performs feature statistics transferring via modulation with per-point attention-weighted mean and variance of the style feature. Nevertheless, AdaAttN will inevitably lose some global style information, since it replaces the global mean and variance with the attention-weighted mean and variance that focus more on local style information. Kwon *et al.* [26] introduced CLIP (Contrastive Language-Image Pre-Training) [36] into the style transfer task and used a text description instead of a style image to represent the desired style. Fu *et al.* [11] presented another text-guided style transfer framework that learns the correlation between text prompts and style images based on large amounts of paired training data. These methods empower users to create more creative artistic images with input texts, yet the stylization results are far from satisfactory in terms of content preservation and style transformation.

Unlike these existing methods, our approach not only takes both global and local style into consideration without involving any semantic information of the style image, but also learns better content and style representations.

3 PROPOSED METHOD

Here, we first describe the overall pipeline of our approach in Section 3.1. Then, we give details of the proposed Two-Stage Statistics-Aware Transformation (TSSAT) module in Section 3.2. Finally, Section 3.3 introduces the loss functions used in our model, including our newly proposed attention-based content loss and patch-based style loss.

3.1 Overview

Formally, our task can be described as follows: given an arbitrary content image I_c and an arbitrary style image I_s , we aim to learn a generative model to synthesize the corresponding stylized image I_{cs} that not only preserves the content structures of I_c but also learns the local and global style patterns from I_s . The overall framework of our approach is illustrated in Figure 2. As we can see, there are mainly three components in our model: an encoder E , a two-stage statistics-aware transformation module $TSSAT$, and a decoder D .

In detail, the encoder E is a pre-trained VGG-19 network [39] ϕ whose parameters are fixed during training. We feed the content image I_c and style image I_s to ϕ to extract their respective VGG feature maps,

$$F_c := \phi(I_c), \quad F_s := \phi(I_s) \quad (1)$$

After obtaining the content features F_c and style features F_s , we employ a two-stage statistics-aware transformation module $TSSAT$ to match the global and local statistics of F_c with those of F_s , yielding F_{gl} as the fusion result of content and style information,

$$F_{gl} := TSSAT(F_c, F_s) \quad (2)$$

Finally, we input F_{gl} to the decoder D to generate the stylized image I_{cs} ,

$$I_{cs} := D(F_{gl}) \quad (3)$$

Besides, it is also worth noting that we use F_{cs} to represent the image features extracted from I_{cs} via the VGG-19 network ϕ , i.e., $F_{cs} := \phi(I_{cs})$.

3.2 Two-Stage Statistics-Aware Transformation

The key idea of TSSAT is to harness feature statistics to first build the global style foundation (in the global statistics alignment stage) and then enrich local style details (in the local statistics swap stage), simulating the drawing process of humans. In this paper, we mainly take mean and variance as the feature statistics to show TSSAT's effectiveness. Its detailed structure is depicted in Figure 2 (dashed grey box).

Global statistics alignment. This stage aims to align the global statistics of content and style features to learn global style. Previous works [15, 21, 31] have demonstrated mean and variance do not carry any semantic information but only the style information of an image. In this way, we can achieve our goal by first normalizing the content features F_c and then scaling and shifting the normalized F_c with the corresponding scalar components of the style features F_s ,

$$F_g := \sigma(F_s) \left(\frac{F_c - \mu(F_c)}{\sigma(F_c)} \right) + \mu(F_s) \quad (4)$$

where σ and μ denote the mean and standard deviation of feature maps, respectively. F_g is the output, which captures the global style distribution of the style image.

Local statistics swap. In this stage, we take F_g as the content features and aim to introduce more local style patterns to the stylization result by swapping the local statistics of content and style features. The detailed procedure is as follows:

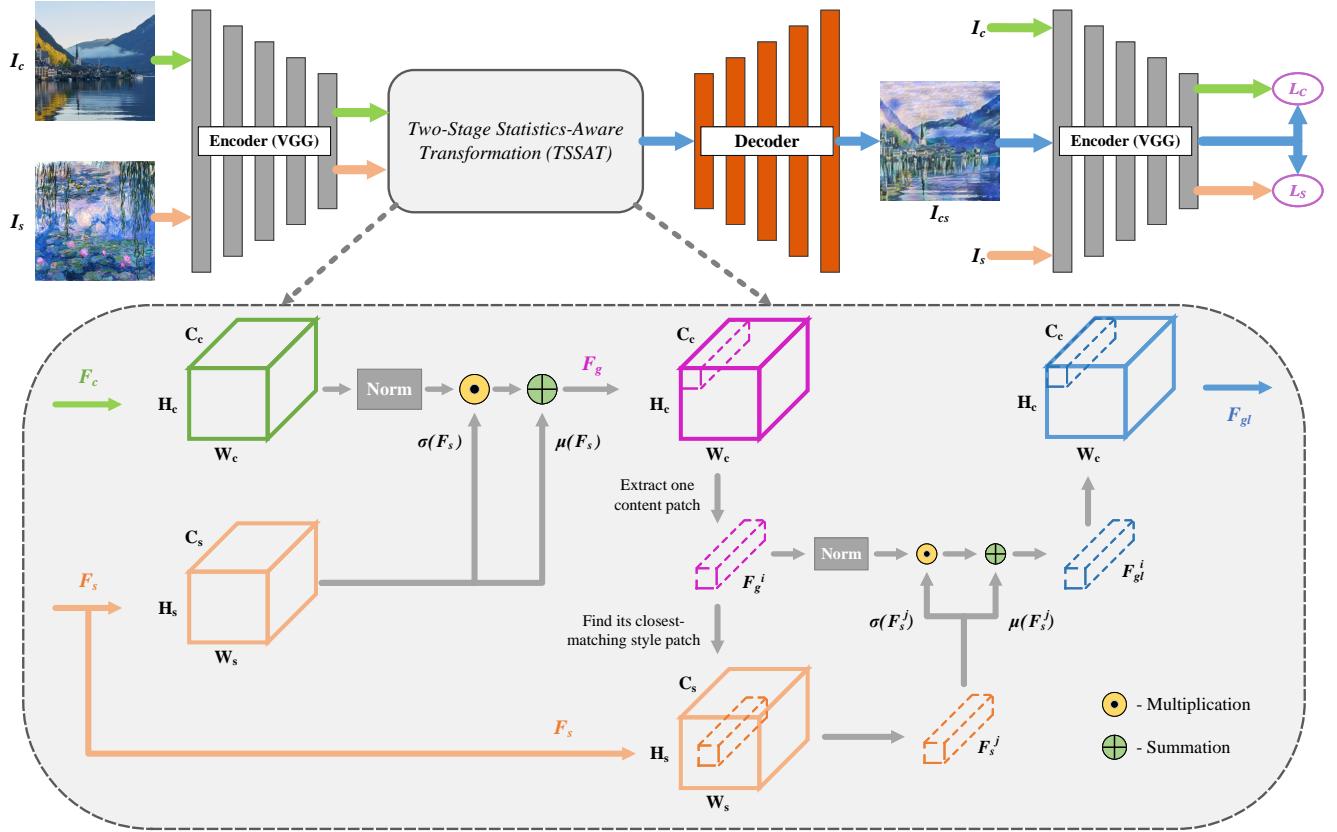


Figure 2: Overview of our framework. (1) We take a fixed pre-trained VGG-19 network as our encoder to extract content and style features. **(2)** The TSSAT module performs style transfer in the feature space by first aligning the global statistics of content and style features and then swapping the local statistics in a patch-wise manner. **(3)** The decoder inverts the deep transferred features into artistic images.

- (1) Extract a set of $k \times k$ patches for both F_g and F_s , denoted by $\{F_g^i\}_{i \in n_g}$ and $\{F_s^j\}_{j \in n_s}$, where n_g and n_s are the number of extracted patches.
 - (2) For each content patch F_g^i , find its closest-matching style patch F_s^j through a convolution layer, where the normalized style feature patches $\{F_s^j / \|F_s^j\|\}_{j \in n_s}$ are the filters and F_g^i is the input. The output of this layer is a vector, where each scalar is equivalent to the cosine similarity between the content patch and one style patch. This way, the closest-matching style patch F_s^j can be found by determine the index of the maximum value in the vector.
 - (3) Swap the mean and variance of the content patch F_g^i with those of its closest-matching style patch F_s^j ,
- $$F_{gl}^i := \sigma(F_s^j) \left(\frac{F_g^i - \mu(F_g^i)}{\sigma(F_g^i)} \right) + \mu(F_s^j) \quad (5)$$
- (4) Recombine the feature patches $\{F_{gl}^i\}_{i \in n_g}$ to obtain the feature maps F_{gl} , where abundant and fine-grained local style patterns are involved on the basis of F_g .

Note that to prevent the semantic information of the style image from being introduced into the stylization result, here we overcome the inertial thinking of previous methods [7, 28, 35, 46, 51] and propose to learn local styles based on feature statistics instead of neural patches, which is an elegant and effective reformation. We also emphasize that since the local statistics swap operation is performed between every two most similar patches of the global-style-aligned features F_g and F_s , it will not result in global style deviation but only further enrich local style details (see detailed demonstrations in Section 4.4).

Thanks to the above two stages, our TSSAT is able to synthesize more appealing stylization results with elaborately decorated style patterns. TSSAT also allows flexible style pattern modulation by adjusting the patch size k in the local statistics swap stage, leading to more diverse stylization results (see detailed demonstrations in Section 4.2).

3.3 Loss Functions

The loss functions used in our model consist of the content loss L_c , the attention-based content loss L_{ac} , the style loss L_s , the patch-based style loss L_{ps} , and the identity loss $L_{identity}$. Among

them, \mathcal{L}_{ac} and \mathcal{L}_{ps} are our newly proposed losses. Details of each loss will be explained in the remaining part of this section.

Content loss. We learn the content information by minimizing the perceptual differences between the content image I_c and the stylized image I_{cs} ,

$$\mathcal{L}_c := \sum_{i=1}^L \|\phi_i(I_c) - \phi_i(I_{cs})\|_2 \quad (6)$$

where ϕ_i denotes the i_{th} layer of VGG-19. We use `relu4_1` and `relu5_1` layers in our experiments.

Attention-based content loss. \mathcal{L}_c focuses on pulling every stylized feature point closer to the corresponding content feature point, neglecting the semantic relation among different feature points within an image. To further enhance the semantic correspondence between I_c and I_{cs} , we first capture the semantic relation within an image based on the self-attention mechanism [8, 52] and then enforce the attention map derived from I_{cs} to be consistent with that derived from I_c . To provide deterministic supervision signals, we use a parameter-free version of attention map without the learnable 1×1 convolution kernels [33],

$$A(x) := \text{Softmax}(\text{Norm}(x)^T \otimes \text{Norm}(x)) \quad (7)$$

where \otimes denotes matrix multiplication. However, we found that the diagonal elements of the resulting attention map are very close to 1 and most of the remaining elements are 0. We argue that this is because each feature point is much more similar to itself than to other feature points and the gap is further greatly magnified by the Softmax operation. To make the attention map focus more on the inter-point relation, we remove the diagonal elements before Softmax. Nevertheless, the attention map is still a sparse matrix since the similarity between neighboring feature points is generally way above average, making other points ignored. To bridge the large gap and take more inter-point relations into account, we further scale down the absolute value of each element in the similarity matrix (before Softmax) by a factor of τ so that the resulting attention map will be a dense matrix. The above process is depicted in Figure 3 and formulated as,

$$A'(x) := \text{Softmax}(\overline{\text{diag}}(\text{Norm}(x)^T \otimes \text{Norm}(x))/\tau) \quad (8)$$

where $\overline{\text{diag}}$ denotes the operation of removing diagonal elements. As a result, the attention-based content loss can be defined as,

$$\mathcal{L}_{ac} := \sum_{i=1}^L \|A'(\phi_i(I_c)) - A'(\phi_i(I_{cs}))\|_2 \quad (9)$$

For ϕ_i , `relu4_1` and `relu5_1` layers are used in our experiments. Note that \mathcal{L}_{ac} is significantly different from the content loss in STROTSS [23], which attempts to maintain the relative pairwise similarities between *some randomly chosen locations* in an image, while the attention map used in our loss considers the semantic relations between *every two different feature points* within an image and thus enables better content preservation. In addition, \mathcal{L}_{ac} is based on the *self-attention mechanism*, which is more effective in capturing semantic relations.

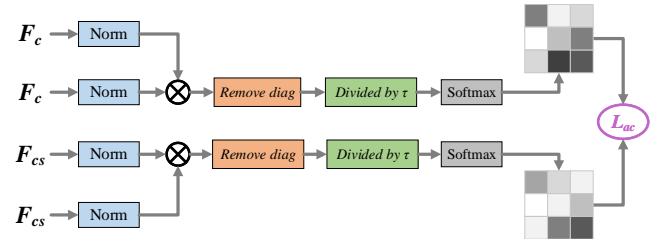


Figure 3: Illustration of our proposed attention-based content loss \mathcal{L}_{ac} .

Style loss. The style loss is calculated by matching the mean and standard deviation of the style features to those of the stylized features,

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

where we use `relu1_1`, `relu2_1`, `relu3_1`, `relu4_1`, and `relu5_1` layers to calculate this loss.

Patch-based style loss. \mathcal{L}_s constrains the similarity between I_s and I_{cs} from a global perspective. To further enhance the stylization effect from a local perspective, we propose a patch-based style loss, which encourages the style similarity between each stylized feature patch and its closest-matching style feature patch via statistics alignment,

$$\mathcal{L}_{ps} := \sum_{i=1}^N \|\mu(\psi_i(\phi_{r4_1}(I_{cs}))) - \mu(\psi_{NN(i)}(\phi_{r4_1}(I_s)))\|_2 + \|\sigma(\psi_i(\phi_{r4_1}(I_{cs}))) - \sigma(\psi_{NN(i)}(\phi_{r4_1}(I_s)))\|_2 \quad (11)$$

where $r4_1$ represents `relu4_1`, ψ_i denotes the i_{th} patch of the stylized features, and $\psi_{NN(i)}$ denotes its best-matching style feature patch. The patch size is the same as that in the local statistics swap operation. It is worth mentioning that the MRF loss [27] also considers the local style similarity between I_s and I_{cs} , yet it calculates the distance between two closest-matching feature patches rather than their statistics, which will inevitably introduce some semantic information from I_s to I_{cs} .

Identity loss. Following [3, 9, 35], we also adopt an identity loss to better maintain the content structure and style characteristics simultaneously,

$$\mathcal{L}_{identity} := \lambda_{id1} (\|I_c - I_{cc}\|_2 + \|I_s - I_{ss}\|_2) + \lambda_{id2} \sum_{i=1}^L (\|\phi_i(I_c) - \phi_i(I_{cc})\|_2 + \|\phi_i(I_s) - \phi_i(I_{ss})\|_2) \quad (12)$$

where I_{cc}/I_{ss} are the generated results when the input images are two identical content/style images. λ_{id1} and λ_{id2} are hyperparameters controlling weights of their corresponding loss terms. The VGG-19 layers used here include `relu1_1`, `relu2_1`, `relu3_1`, `relu4_1`, and `relu5_1`.

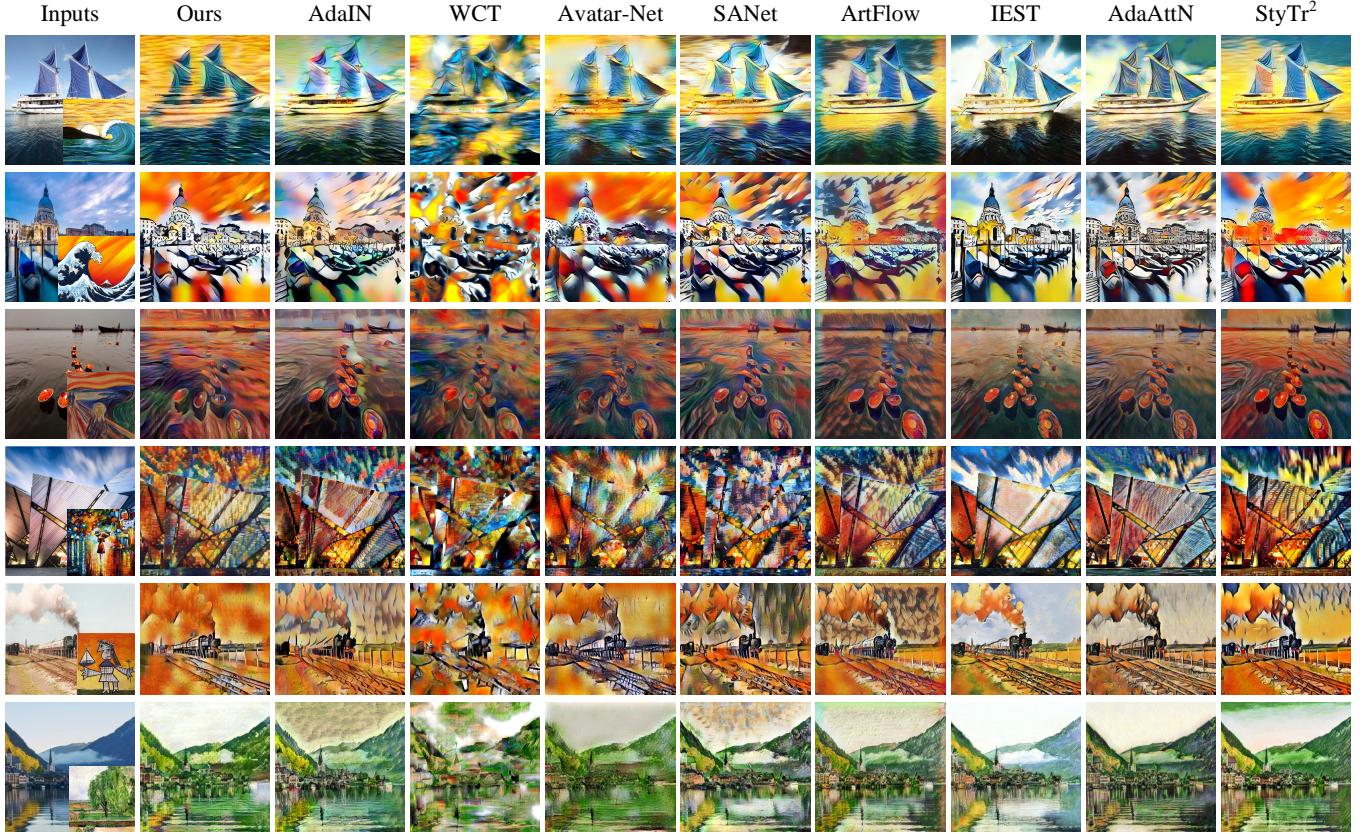


Figure 4: Qualitative comparisons. The first column shows the input content and style images. The rest of the columns show the stylization results generated with different style transfer methods. Please zoom in to compare the details.

Final objective. We summarize all aforementioned losses and obtain the final objective of our model,

$$\mathcal{L} := \lambda_1 \mathcal{L}_c + \lambda_2 \mathcal{L}_{ac} + \lambda_3 \mathcal{L}_s + \lambda_4 \mathcal{L}_{ps} + \lambda_5 \mathcal{L}_{identity} \quad (13)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, and λ_5 are the balancing weights for different loss terms.

4 EXPERIMENTS

This section is organized as follows: Section 4.1 introduces the implementation details, datasets, and baselines. Section 4.2 and Section 4.3 present the qualitative and quantitative results, respectively. Finally, the effect of each component in our model is explored in Section 4.4.

4.1 Experimental Settings

Implementation details. As introduced in Section 3, our model mainly consists of three components: an encoder E, a decoder D, and a two-stage statistics-aware transformation module TSSAT. Among them, E is a fixed pre-trained VGG-19 network (up to *relu4_1*) [39], and D is symmetrical to E. To be more specific, all pooling layers in E are replaced by nearest upsampling to form D. As for TSSAT, its architecture has been illustrated in Figure 2 (dashed grey box)

and its input features are extracted from the *relu4_1* layer of VGG-19. The patch size k in the local statistics swap operation is set to 5 during training, and different patch sizes can be employed to adjust local style patterns at inference. If not specifically stated, our stylization results in this paper are generated when $k = 5$ by default. The hyper-parameter τ in Equation (8) is set to 100. The loss weights in Equation (12) and (13) are set to $\lambda_{id1} = 50$, $\lambda_{id2} = 1$, $\lambda_1 = 5$, $\lambda_2 = 50000$, $\lambda_3 = 6$, $\lambda_4 = 0.5$, and $\lambda_5 = 1$. We train our network using the Adam optimizer [22] with a learning rate of 0.0001 and a batch size of 4 for 160000 iterations. Our code is available at <https://github.com/HalbertCH/TSSAT>.

Datasets. We take MS-COCO [32] and WikiArt [20] as our content dataset and style dataset, respectively. During the training stage, we initially resize the smallest dimension of training images to 512 while maintaining the aspect ratio. Subsequently, we randomly crop patches of size 256×256 from these images to serve as input. In the reference stage, our method is capable of handling content and style images of any size.

Baselines. We select 8 state-of-the-art style transfer methods to compare with our approach, including AdaIN [15], WCT [30], Avatar-Net [38], SANet [35], ArtFlow [1], IEST [3], AdaAttN [33], and StyTr² [9]. For all the above baselines, we use their publicly available implementations to produce the results.

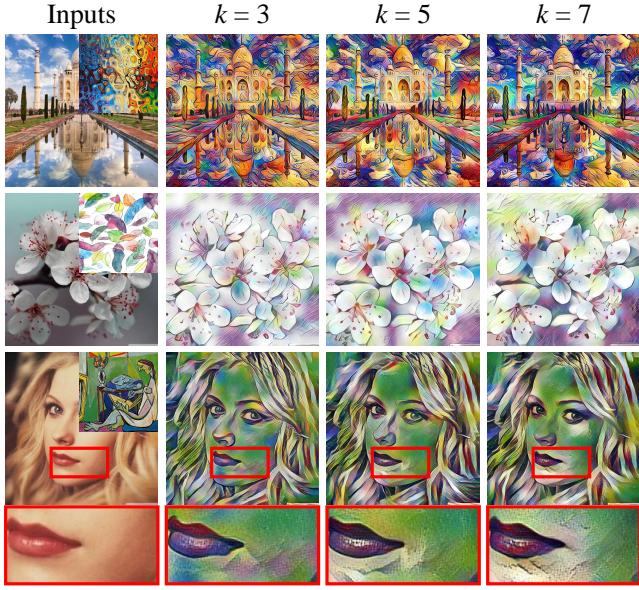


Figure 5: Stylization results with different patch sizes k .

4.2 Qualitative Results

We present qualitative stylization results of different style transfer methods in Figure 4 for comparison. It can be observed that AdaIN [15] often captures insufficient style patterns and introduces some abrupt colors that do not exist in the style image (e.g., 1st, 2nd, and 6th rows). WCT [30] has severe problems with content preservation (e.g., 2nd, 4th, and 6th rows). Avatar-Net [38] suffers from the content structure blur and style pattern distortion issues (e.g., 1st, 2nd, and 4th rows). SANet [35] sometimes introduces undesired semantic structures from the style image to the stylization result (e.g., 3rd, 4th, and 5th rows). ArtFlow [1] tends to produce unwanted artifacts in relatively smooth regions (e.g., 2nd, 3rd, and 5th rows). The results of IEST [3] are generally less stylized with limited colors and textures (e.g., 2nd, 4th, and 5th rows). For AdaAttN [33], there is an obvious style deviation between the style image and the stylized image generated by it (e.g., 1st, 5th, and 6th rows). The results of StyTr² [9] usually have the problem of color oversaturation, resulting in inconsistent colors with the style image (e.g., 1st, 2nd, and 3rd rows). In comparison, our method TSSAT not only captures accurate and adequate style patterns, but also retains clear and clean content structures, as shown in the 2nd column of Figure 4.

Please zoom in to compare the details.

As introduced in Section 3.2, our proposed TSSAT also allows flexible style pattern modulation by adjusting the patch size k in the local statistics swap stage. Note that our model can adapt to different patch sizes at inference once trained with one patch size. Therefore, it is very convenient and efficient for our model to produce diverse stylization results with different patch sizes, as shown in Figure 5. We can see that the stylized images are more colorful and vivid when the patch size is small and become cleaner and neater when the patch size is bigger. The zoom-in regions demonstrate the change of local style patterns more clearly.

4.3 Quantitative Results

The qualitative results presented above could be subjective. In this section, we adopt several quantitative metrics to conduct more comprehensive and objective evaluations.

Perceptual distance and GELP. Perceptual distance [12, 18] estimates the multi-level feature distances between the content and stylized images. It is usually taken as the content loss by existing style transfer methods, including our and competing methods. Following [1, 9, 47], here we adopt it to measure the performance of content preservation. Meanwhile, to measure the performance of style transformation, we employ the GELP metric [41, 43] that takes both global style effects (including global colors and holistic textures) and local style patterns (including the similarity and diversity of the local style patterns) into consideration. We randomly select 50 content-style pairs for each method and report the average perceptual distance and GELP score in Table 1. As we can see, our proposed TSSAT obtains the lowest perceptual distance and the third-highest GELP score. The results indicate that our method achieves the best trade-off between content preservation and style transformation, which is consistent with the visual comparisons in Figure 4. We also compare the performance of our method under different patch sizes k . It is easy to find that: the bigger the patch size, the better the content preservation; the smaller the patch size, the better the style transformation.

Preference score. We further perform a user study [1, 2, 4, 35, 45, 50, 54] to investigate user preference over different stylization results. Specifically, we first choose 15 content images and 10 style images to form 150 content-style pairs. Then, we randomly sample 20 content-style pairs for each subject and synthesize 9 different stylized images for each pair using 9 style transfer methods (including our method and 8 baselines). Next, we ask the subject to indicate his/her favorite stylization result for each content-style pair. Finally, we collect 1000 votes from 50 subjects and show the percentage of votes for each method in Table 1, where we can observe that our method achieves preferable performance than competitors by a significant margin.

Efficiency analyses. We compare the efficiency of our method with prior works in the bottom row of Table 1. All the methods are tested on a single Nvidia GeForce RTX 3090 GPU with the image size of 512×512. It can be observed that the speed of our method is comparable with the state-of-the-art methods such as ArtFlow [1] and StyTr² [9]. Moreover, the speed of our method can be further accelerated by increasing the patch size k .

4.4 Ablation Study

Analyses of the TSSAT module. As introduced in Section 3.2, the TSSAT module consists of a global statistics alignment (*abbr.* GSA) stage and a local statistics swap (*abbr.* LSS) stage. To study their effects, we compare the stylization results of our method with and without GSA/LSS in Figure 6 (c-e). We can see that without GSA, the model neglects the global style distribution of the style image. Without LSS, the model fails to capture abundant and fine-grained local colors and texture patterns. This is because GSA and LSS are responsible for global style learning and local style capturing, respectively. We can also see that the LSS stage performed after GSA influences the global style only to a restricted extent and

Table 1: Quantitative comparisons. Dis. stands for distance. We show the best results in bold, the second-best results with a star*, and the third-best results with an underline.

	AdaIN	WCT	Avatar-Net	SANet	ArtFlow	IEST	AdaAttN	StyTr ²	Ours		
									$k = 3$	$k = 5$	$k = 7$
Perceptual Dis. \downarrow	2.061	2.828	2.297	2.251	2.052	1.876*	2.191	1.958	2.019	<u>1.880</u>	1.778
GELP Score \uparrow	1.446	1.457	1.488*	1.511	1.451	1.382	1.438	1.461	<u>1.481</u>	1.476	1.469
Preference (%) \uparrow	0.059	0.053	0.061	0.094	0.070	<u>0.152</u>	0.105	0.174*	-	0.232	-
Time (sec) \downarrow	0.062	0.997	0.308	<u>0.077</u>	0.341	0.074*	0.112	0.401	0.484	0.337	0.329

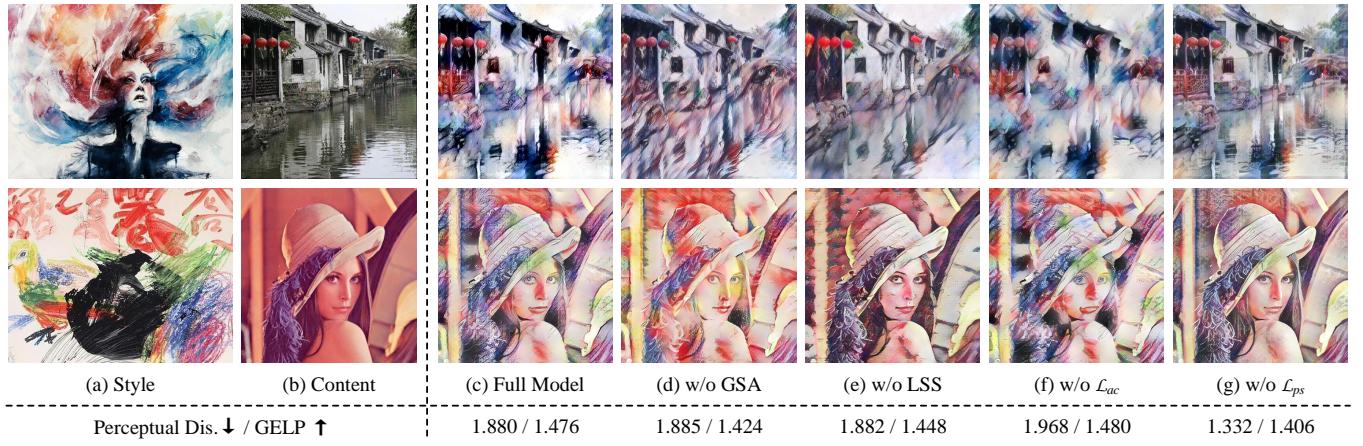


Figure 6: Ablation study results. The first two columns show the style and content images, respectively. The rest columns show the stylization results generated by our model under different settings.

will not result in global style deviation. We can get an explanation from the relation between the two stages: GSA builds the global style foundation and LSS just further enriches local style details based on the foundation. Above analyses are also supported by the quantitative results reported in the last row.

Loss analyses. To investigate the influence of the attention-base content loss \mathcal{L}_{ac} and the patch-based style loss \mathcal{L}_{ps} , we remove them from our model and show the experimental results in Figure 6 (f) and (g). It can be observed that without \mathcal{L}_{ac} , the content structures of the stylized image become less clear, and notable distortions are introduced. The results demonstrate the importance of \mathcal{L}_{ac} in content preservation. In addition, we also find that without \mathcal{L}_{ps} , the stylization results become less colorful and vivid, and lots of local style information is lost. It indicates that \mathcal{L}_{ps} is of great significance in local style learning. The qualitative results, together with the quantitative results reported in the last row, verify that only the full model can achieve satisfying performance in both content preservation and style transformation.

5 CONCLUSION AND LIMITATION

In this paper, we propose a Two-Stage Statistics-Aware Transformation (TSSAT) module and two loss functions to improve the style transformation and content preservation effect of artistic style transfer. The contribution of TSSAT is the idea of harnessing feature

statistics to first build the global style foundation (in the global statistics alignment stage) and then further enrich local style details (in the local statistics swap stage), simulating the drawing process of humans. The feature statistics we adopt in this paper are mean and variance, and more alternatives can be explored in the future. The attention-based content loss enables better content preservation by enforcing the semantic relation in the content image to be retained during stylization. The patch-based style loss facilitates local style learning by encouraging the similarity between each stylized feature patch and its closest-matching style feature patch. Extensive experiments demonstrate the effectiveness and superiority of our proposed method.

A main limitation of this work is that the proposed method is not fast enough to achieve real-time style transfer. This is because the local statistics swap operation in our TSSAT module needs to be conducted for many times between different feature patches, which is kind of time consuming (the smaller the patch size, the slower the speed). We will take the efficiency issue as our future work and try to simplify the local statistics swap operation for higher execution speed.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation of China (Grant Nos. 62072242).

REFERENCES

- [1] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. 2021. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 862–871.
- [2] Prashant Chandran, Gaspard Zoss, Paulo Gotardo, Markus Gross, and Derek Bradley. 2021. Adaptive convolutions for structure-aware style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7972–7981.
- [3] Haibo Chen, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, Dongming Lu, et al. 2021. Artistic Style Transfer with Internal-external Learning and Contrastive Learning. *Advances in Neural Information Processing Systems* 34 (2021).
- [4] Haibo Chen, Lei Zhao, Lihong Qiu, Zhizhong Wang, Huiming Zhang, Wei Xing, and Dongming Lu. 2020. Creative and diverse artwork generation using adversarial networks. *IET Computer Vision* 14, 8 (2020), 650–657.
- [5] Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2021. DualAST: Dual Style-Learning Networks for Artistic Style Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 872–881.
- [6] Haibo Chen, Lei Zhao, Huiming Zhang, Zhizhong Wang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2021. Diverse image style transfer via invertible cross-space mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14880–14889.
- [7] Tian Qi Chen and Mark Schmidt. 2016. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337* (2016).
- [8] Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long Short-Term Memory-Networks for Machine Reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 551–561.
- [9] Yingyong Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. 2022. StyTr2: Image Style Transfer with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11326–11336.
- [10] Yingyong Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. 2020. Arbitrary style transfer via multi-adaptation network. In *Proceedings of the 28th ACM international conference on multimedia*. 2719–2727.
- [11] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. 2022. Language-driven artistic style transfer. In *European Conference on Computer Vision*. Springer, 717–734.
- [12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2414–2423.
- [13] Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. 2019. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning*. PMLR, 2722–2730.
- [14] Zhiyuan Hu, Jia Jia, Bei Liu, Yaohua Bu, and Jianlong Fu. 2020. Aesthetic-aware image style transfer. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3320–3329.
- [15] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.
- [16] Jing Huo, Shiyin Jin, Wenbin Li, Jing Wu, Yu-Kun Lai, Yinghuan Shi, and Yang Gao. 2021. Manifold Alignment for Semantically Aligned Style Transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14861–14869.
- [17] Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. 2020. Dynamic instance normalization for arbitrary style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4369–4376.
- [18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.
- [19] Nikolai Kalischeck, Jan D Wegner, and Konrad Schindler. 2021. In the light of feature distributions: moment matching for Neural Style Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9382–9391.
- [20] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. 2013. Recognizing image style. *arXiv preprint arXiv:1311.3715* (2013).
- [21] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.
- [22] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [23] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. 2019. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10051–10060.
- [24] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. 2019. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*. 4422–4431.
- [25] Dmytro Kotovenko, Artsiom Sanakoyeu, Pingchuan Ma, Sabine Lang, and Bjorn Ommer. 2019. A content transformation block for image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10032–10041.
- [26] Gihyun Kwon and Jong Chul Ye. 2022. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18062–18071.
- [27] Chuan Li and Michael Wand. 2016. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2479–2486.
- [28] Jie Li, Liwen Wu, Dan Xu, and Shaowen Yao. 2022. Arbitrary style transfer with attentional networks via unbalanced optimal transport. *IET Image Processing* 16, 7 (2022), 1778–1792.
- [29] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. 2019. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3809–3817.
- [30] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal style transfer via feature transforms. In *Advances in neural information processing systems*. 386–396.
- [31] Tianwei Lin, Zhuoqi Ma, Fu Li, Dongliang He, Xin Li, Errui Ding, Nannan Wang, Jie Li, and Xinbo Gao. 2021. Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5141–5150.
- [32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
- [33] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. 2021. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6649–6658.
- [34] Shiguang Liu and Ting Zhu. 2021. Structure-guided arbitrary style transfer for artistic image and video. *IEEE Transactions on Multimedia* 24 (2021), 1299–1312.
- [35] Dae Young Park and Kwang Hee Lee. 2019. Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5880–5888.
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.
- [37] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Bjorn Ommer. 2018. A style-aware content loss for real-time hd style transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 698–714.
- [38] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. 2018. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8242–8250.
- [39] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [41] Zhizhong Wang, Zhanjie Zhang, Lei Zhao, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2022. AesUST: Towards Aesthetic-Enhanced Universal Style Transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1095–1106.
- [42] Zhizhong Wang, Lei Zhao, Haibo Chen, Lihong Qiu, Qihang Mo, Sihu Lin, Wei Xing, and Dongming Lu. 2020. Diversified arbitrary style transfer via deep feature perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7789–7798.
- [43] Zhizhong Wang, Lei Zhao, Haibo Chen, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2021. Evaluate and improve the quality of neural style transfer. *Computer Vision and Image Understanding* 207 (2021), 103203.
- [44] Zhizhong Wang, Lei Zhao, Haibo Chen, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2022. DivSwapper: Towards Diversified Patch-based Arbitrary Style Transfer. In *Proc. Int. Joint Conf. on Artif. Intell. (IJCAI)*. 4980–4987.
- [45] Zhizhong Wang, Lei Zhao, Zhiwen Zuo, Ailin Li, Haibo Chen, Wei Xing, and Dongming Lu. 2023. MicroAST: Towards super-fast ultra-resolution arbitrary style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 2742–2750.
- [46] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. 2021. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14618–14627.
- [47] Zijie Wu, Chunjin Song, Yang Zhou, Minglun Gong, and Hui Huang. 2020. Efafnet: Exchangeable feature alignment network for arbitrary style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12305–12312.
- [48] Zijie Wu, Zhen Zhu, Junping Du, and Xiang Bai. 2022. CCPL: Contrastive Coherence Preserving Loss for Versatile Style Transfer. In *European Conference on Computer Vision*. Springer, 189–206.

- [49] Wenju Xu, Chengjiang Long, Ruisheng Wang, and Guanghui Wang. 2021. Drbgan: A dynamic resblock generative adversarial network for artistic style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6383–6392.
- [50] Jinchao Yang, Fei Guo, Shuo Chen, Jun Li, and Jian Yang. 2022. Industrial Style Transfer with Large-scale Geometric Warping and Content Preservation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7834–7843.
- [51] Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. 2019. Attention-aware multi-stroke style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1467–1475.
- [52] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-attention generative adversarial networks. In *International conference on machine learning*. PMLR, 7354–7363.
- [53] Yulun Zhang, Chen Fang, Yilin Wang, Zhaowen Wang, Zhe Lin, Yun Fu, and Jimei Yang. 2019. Multimodal style transfer via graph cuts. In *Proceedings of the IEEE International Conference on Computer Vision*. 5943–5951.
- [54] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. 2022. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8035–8045.
- [55] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. 2022. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–8.
- [56] Zhiwen Zuo, Lei Zhao, Shuobin Lian, Haibo Chen, Zhizhong Wang, Ailin Li, Wei Xing, and Dongming Lu. 2022. Style fader generative adversarial networks for style degree controllable artistic style transfer. In *Proc. Int. Joint Conf. on Artif. Intell. (IJCAI)*. 5002–5009.