Learning Dynamic Style Kernels for Artistic Style Transfer

Wenju Xu
OPPO US Research Center
Palo Alto, CA, USA
xuwenju123@gmail.com

Chengjiang Long
Meta Reality Lab
Burlingame, CA, USA
clong1@meta.com

Yongwei Nie
South China University of Technology
Guangzhou, Guangdong, China
nieyongwei@scut.edu.cn

Figure 1. The first two rows show comparisons with state-of-the-art methods. The last row shows results of our proposed method, which can faithfully transfer various styles and preserve the content structure.

Abstract

Arbitrary style transfer has been demonstrated to be efficient in artistic image generation. Previous methods either globally modulate the content feature ignoring local details, or overly focus on the local structure details leading to style leakage. In contrast to the literature, we propose a new scheme "style kernel" that learns spatially adaptive kernels *for per-pixel stylization, where the convolutional kernels are dynamically generated from the global style-content aligned feature and then the learned kernels are applied to modulate the content feature at each spatial position. This new scheme allows flexible both global and local interactions between the content and style features such that the wanted styles can be easily transferred to the content image while at the same time the content structure can be easily preserved. To further enhance the flexibility of our style transfer method, we propose a Style Alignment Encoding (SAE) module complemented with a Content-based Gating Modulation (CGM) module for learning the dynamic style kernels in focusing regions. Extensive experiments strongly demonstrate that our proposed method outperforms state-of-the-art methods and exhibits superior performance in terms of visual quality and efficiency.*

1. Introduction

Artistic style transfer [48] refers to a hot computer vision technology that allows us to recompose the content of an image in the style of an artistic work. Figure 1 shows several vivid examples. We might have ever imagined what a photo might look like if it were painted by a famous artist like Pablo Picasso or Van Gogh. Now style transfer is the computer vision technique that turns this into a reality. It has great potential values in various real-world applications and therefore attracts a lot of researchers to constantly put efforts to make progress towards both quality and efficiency.

Most of existing style transfer works [6, 15, 18, 26, 33] either globally modulate the content feature ignoring local details or overly focus on the local structure details leading to style leakage. In particular, [15, 18] seeks to match global statistics between content and style images, resulting in inconsistent stylizations that either remain large parts of the content unchanged or contain local content with distorted style patterns. SANet [33] and AdaAttN [31] improve the content similarity by learning attention maps that match the semantics between the style and stylized images, but these methods tend to distort object instances when improper style patterns are involved. Recently, IEC [5] adopts contrastive learning to pull close both the content and style representations between input and output images. MAST [16]

seeks to balance the style and content via manifold alignment. Although both IEC and MAST have achieved some success in most cases, they are still far from satisfactory to well balance two important requirements, *i.e*., style consistency and structure similarity.

In this paper, we develop a novel style transfer framework in a manner of encoder-decoder structure, as illustrated in Figure 2, which contains two important modules: (1) a Style Alignment Encoding (SAE) module enhanced by Content-based Gating Modulation (CGM), which is used to generate content-style aligned features, and (2) a Style Kernel Generation (SKG) module used to transform the output features of SAE to convolutional kernels. In such a framework, we propose a new style transfer scheme, *"style kernel"*, which treats the features of the style images as dynamic convolutional kernels and then transfer the style information to the content image by convolving the content image by the learned dynamic style kernels. This allows fine-grained local interactions between the style and content features, the flexibility of which makes both style transferring and content structure preserving much easier.

To enforce the global correlation between the content and style features, we simulate the self-attention mechanism of Transformer [41] in the SAE module. This treats the content and style features as the queries and keys of the self-attention operator respectively and computes a large attention map the elements of which measure for each local content context the similarity to each local style context. With the attention map, we can aggregate for each pixel of the content image all the style features in a weighted manner to obtain content-style aligned features. We also design a Content-based Gating Modulation (CGM) operator that further thresholds the attention map and zeros the places where similarities are too small, seeking an adaptive number of correlated neighbors as a focusing region for each query point. Then, we use another set of convolutions that constitute the SKG module to further transform the content-style aligned features, and view (reshape) the output of SKG as dynamic style kernels. In order to improve the efficiency of the dynamic convolutions, SKG predicts separable local filters (two 1D filters and a bias) instead of a spatial 2D filter that has more parameters. The learned style kernels already integrate the information of the given content and style images. We further apply the learned dynamic style kernels to the content image feature for transferring the target style to the content image.

We shall emphasize that unlike *"style code"* in AdaIN [15], DRB-GAN [48], and AdaConv [2] modeled as the dynamic parameters (*e.g*. mean, variance, convolution kernel) which are shared over all spatial positions globally without distinguishing the semantic regions, the learned dynamic kernels via our *"style kernel"* are point-wisely modeled upon the globally style-content aligned feature, and therefore is able to make full use of the point-wise semantic structure correlation to modulate content feature for better artistic style transfer. Our learned dynamic style kernels are good at transferring the globally aggregated and locally semantic-aligned style features to local regions. This ensures our style transfer model that works well on consistent stylization with well preserved structure, overcoming the shortages of existing style transfer methods which either globally modulate the content feature ignoring local details (*e.g*. AdaIN, WCT [28]) or overly focus on the local structure details leading to style leakage (*e.g*. AdaAttN [31] and MAST [16]).

The main contributions of this work are 3-fold as follows:

- We are the first to propose the *"style kernel"* scheme for artistic style transfer, which converts the globally style-content alignment features into position point-wisely dynamic convolutional kernels to convolve the features of content images together for style transfer.
- We design a novel architecture, *i.e*. SAE and CGM, to generate the dynamic style kernels by employing the correlation between the content and style image adaptively.
- Extensive experiments demonstrate that our method produces visually plausible stylization results with fine-grained style details and coherence content with the input content images.

2. Related Works

Artistic Style Transfer was initially solved by Gatys *et al*. [12] a pre-trained neural network to synthesize stylizations in an iterative optimization manner, and then a bunch of neural style transfer methods [3,5,11,18,31,38,39] were developed. In particular, AdaIN [15] introduces the adaptive instance normalization to globally match the statistics between the content and style features. For the same purpose, WCT [28] utilizes the whitening and coloring transforms, and LST [27] introduces a linear transformation matrix predicted on content and style features. [7, 22, 42, 51] seek different ways to match the feature distribution. SANet [33] seeks to align the semantics between the content and style features via a self-attention network. StrTr^2 [8] resorts to the transformers [41, 47]. To further preserve the content details, AdaAttN [31] perform attentive normalization on per-point basis, IEC [5] employs stylization relations with contrastive learning and MAST [16] introduces manifold alignment to balance the style and content. However, emphasizing on content details prevents to transfer adequate style information to some degree, and results in style leakage. Recently, dynamic network based style transfer methods [3, 15, 18, 38] extend the model generalization. As a generic extension of AdaIN, AdaConv [2] introduces a parameter network that predicts network parameters used by

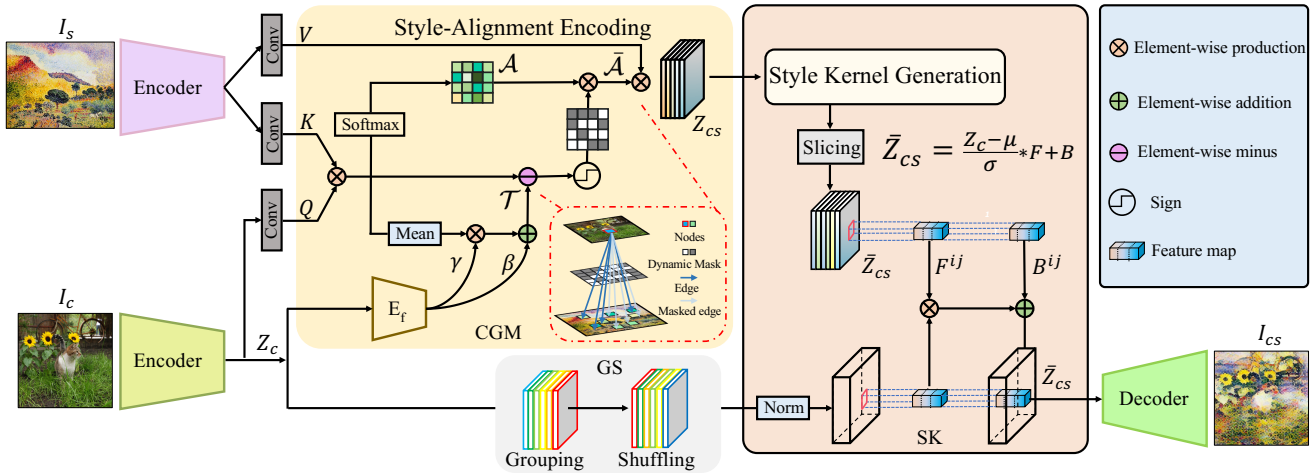

Figure 2. Overview of our proposed scheme *"style kernel"*. A pretrained VGG encoder takes in the content and style image as input to produce the content feature Z_c and style feature Z_s. The style-alignment encoding model employs an attention mechanism by taking the Z_c as query and Z_s as key and value to exploit their semantics matching. The content-based gating module (CGM) learns a dynamic threshold and adaptively adjusts attention weights for a globally content-style aligned feature Z_{cs}, which is transformed by style kernel generation module to dynamic style kernels. Finally, we conduct convolution operations on the channel-wisely grouped and shuffled (GS) content feature with the learned dynamic style kernels to get $\bar{Z}_{cs}$ fed into the decoder to produce the stylizations.

another network for style transfer. DRB-GAN [48] models *"style code"* as the global mean and variance for dynamic convolutions to conduct style transfer. Different from all the above methods, we propose a new scheme *"style kernel"* to learn dynamic convolutional kernels from the style-content alignment feature to point-wisely modulate the content feature for artistic style transfer.

Self-Attention Mechanism is introduced in Transformer [41] as a new attention mechanism. Similar to non-local neural networks [13, 21, 41, 45], Transformer directly works on sequences of image patches to aggregate information for long-range dependencies. The recent researches demonstrate it has been successfully applied in various vision tasks like image recognition [4, 10, 14, 29], object detection [1, 44, 46], image captioning [9], image enhancement [49, 50], and text conditioned image generation [35, 37]. Our model takes the self-attention mechanism to align the semantics of the style feature to that of the content image.

Dynamic Network consists of convolution layers for which the filter parameters are predicted by another sub-network. Recently, there is rapid progress in vision applications [17, 32, 36, 48, 52] that benefits from dynamic network for image/video enhancement. [19] employs dynamic upsampling filters for high-resolution image reconstruction. [52] predicts filters from spatio-temporal features for video deblurring. Differently, our StyleKernel learns spatially adaptive kernel for per-pixel stylization.

3. Our Method

While existing dynamic kernel based methods [2, 15, 48] can efficiently generate stylizations, we observe that their visual-quality is limited in two aspects. (1) As the dynamic kernels are predicted based on the style feature without distinguishing the content-style correlation, the transferred style only globally reflects the style characters of the style image, ruining local content details. (2) The completeness of semantic regions is not well preserved. This is because the dynamic kernels used by the layers are spatially shared.

To alleviate these problems, we propose to align the semantics of the style image to those of the content image and transform the content features for per-pixel stylization in a separable convolution manner. The pipeline of our proposed model is shown in Figure 2. The content image I_c and style image I_s are first fed into a fixed *VGG* encoder to produce the content and style feature maps $Z_c \in \mathbb{R}^{H_c \times W_c \times C_{in}}$ and $Z_s \in \mathbb{R}^{H_s \times W_s \times C_{in}}$, where H, W and C_{in} represent the height, width and channel size of the feature map, respectively. Then in SAE, we align the content and style features by using the content features Z_c as the queries and the style features Z_s as the keys in a self-attention mechanism. The CGM module is conducted on the attention map to seek a focusing region for each query point. Thereafter, we predict by SKG the dynamic style kernels F from the content-style aligned features Z_{cs} . Finally, we group and shuffle (GS) Z_c to break down the correlation within the channel-wise feature and apply convolutions parameterized by the learned kernels to the grouped and shuffled Z_c to obtain $\bar{Z}_{cs}$ which is further decoded to the target stylized image by a decoder.

3.1. Style Alignment Encoding

This module takes the content and style features as input, aligns between them by learning content-style alignment attention, and finally attentively adjusts the attention map by

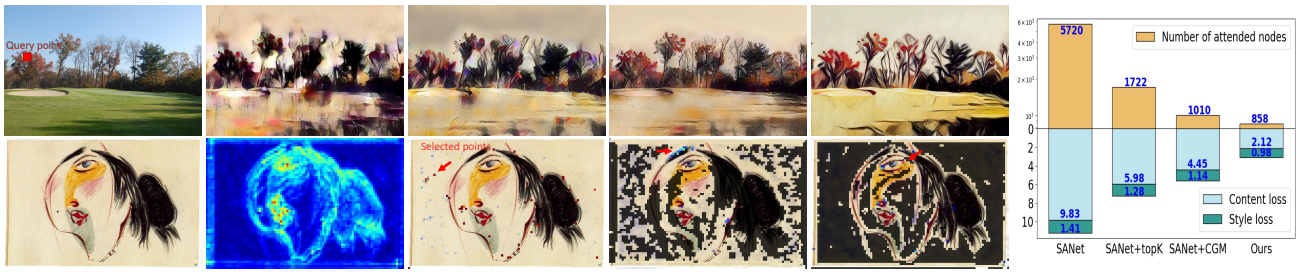

Figure 3. The impact of the feature aggregation. Left: We compare the performances of models conducting different feature aggregation strategies. The SANet takes an attention module to aggregate information across all the nodes. The SANet + topK only acquires information from the nodes with top K attention scores. SANet + CGM takes our content-based gating modulation to aggregate information from a dynamic number of nodes. Our CGM is learned to masked out nodes with low attention score, leading to stylizations with well preserved structure content. The second row shows the distributions of selected nodes (highlighted in colors) and the focused areas (masked-out by black rectangles) learned by CGM. Right: we demonstrate the relation between the content/style loss and the number of selected nodes.

a content-based gating modulation module.

Content-Style Alignment Attention. By taking the content feature Z_c as the query, and the style feature Z_s as the key and value, we compute the content-style alignment attention that seeks to aggregate information from the style feature map according to the matched content information. This warps the style feature to align with the content feature. To this end, we calculate the pairwise feature correlation $\mathcal{M}^{s\to c} \in \mathbb{R}^{H_c W_c \times H_s W_s}$. This results in an attention map $\mathcal{A} \in \mathbb{R}^{H_c W_c \times H_s W_s}$ by:

$$\mathcal{A}(u,v) = \texttt{softmax}(\alpha \mathcal{M}^{s\to c}) = \texttt{softmax}(\alpha \frac{Z_c(u)^T Z_s(v)}{||Z_c(u)||||Z_s(v)||}), \tag{1}$$

where α is the coefficient that controls the sharpness of the $\texttt{softmax}$. $Z_c(u)$ and $Z_s(v)$ stands for the feature of Z_c and Z_s at position $u \in \mathbb{R}^{H_c W_c}$ and $v \in \mathbb{R}^{H_s W_s}$. The u-th row of the attention map $\mathcal{A}$ represents similarities between the u-th node of the content feature and all the nodes of the style feature.

To align the semantics of the style feature to that of the content feature, we aggregate information in Z_s based on the attention score map $\mathcal{A}$ and obtain $Z_{cs} \in \mathcal{R}^{H_c W_c \times C}$ by calculating their weighted average:

$$Z_{cs}(u) = \sum_v \mathcal{A}(u,v) \cdot Z_s(v). \tag{2}$$

where $Z_{cs}(u)$ represents the feature at position u of Z_{cs}. In the following, we denote the $Z_{cs} \in \mathcal{R}^{H_c \times W_c \times C}$, which is obtained by reshaping.

In Eq. 2, all style features are taken into consideration when computing the content-style aligned features Z_{cs}. We find that this yields inferior results because many irrelevant features are involved into the computation. We therefore introduce the Content-based Gating Modulation (CGM) operator which creates dynamic thresholds that further filter out irrelevant style features.

Content-based Gating Modulation is used to further adapt the attention map by the content feature Z_c. Taking Z_c as input, we employ a convolution network E_f to generate a scale parameter λ and a bias parameter β. We then perform a row average operator on the attention matrix $\mathcal{A}$ by:

$$\mathcal{A}_r(i) = \frac{1}{H_c} \sum_{j=1}^{H_c} \mathcal{A}_{i,j}, \forall i \in [1, W_c]. \tag{3}$$

where $\mathcal{A}_r$ is the obtained column vector after the average operator. Then, we obtain the threshold matrix $\mathcal{T}$ by the following equation:

$$\mathcal{T} = \mathcal{A}\left(\lambda \mathcal{A}_r + \beta\right). \tag{4}$$

Finally, we update the attention $\mathcal{A}$ by:

$$\bar{\mathcal{A}} = \mathcal{A} \cdot \texttt{Sign}(\mathcal{A} - \mathcal{T}), \tag{5}$$

where the dot indicates element-wise multiplication and $\texttt{Sign}$ is a function returning 1 given positive input and 0 otherwise. We use $\bar{\mathcal{A}}$ instead of $\mathcal{A}$ in Eq. 2 to compute Z_{cs}.

$\bar{\mathcal{A}}$ is a dynamically learned mask, with which our model creates focused areas and filter out irrelevant style features. This allows our model to aggregate features from fairly correlated features, which is essential to preserve the completeness of semantic regions. In Figure 3 we compare the intermediate visualizations and stylized results. We see that the CGM works to recognize focused areas, where the style information is aggregated by the query point. By contrast, the attention learned by the SANet is distributed over the entire image, including visually less correlated points, such as corners and background.

3.2. Style Kernel Generation

We propose the Style Kernel (SK) Generation network to further transform the content-style aligned features of SAE to dynamic style kernels. It takes the globally content-style

Table 1. Average inference time, measured on a Tesla V100 GPU, for different methods with a batch size of 1 and an input image of 256×256 and 512×512.

Method	AdaIN'17	LST'19	AvatarNet'18	SANet'19	IEC'21	DRB-GAN'21	AdaAttN'21	MAST'21	AdaConv'21	StrTr^2'22	Ours
$T_{256} \downarrow$	0.004	0.004	0.873	0.011	0.012	0.005	0.024	0.126	0.019	0.034	0.017
$T_{256} \downarrow$	0.005	0.005	0.962	0.037	0.021	0.006	0.042	0.239	0.036	0.086	0.035

aligned feature Z_{cs} as input to predict the dynamic style filters $F, B = \{F_1^{ij}, F_2^{ij}, B^{ij}\}$.

Given the content-style aligned feature Z_{cs}, we take two convolution layers to predict the filters $F \in \mathbb{R}^{H_c \times W_c \times C \times 2k}$, which has two 1-dim filters $F_1 = \{F_1^{ij}\}$ and $F_2 = \{F_2^{ij}\}$ of the sizes $k \times 1$ and $1 \times k$, respectively, and one bias vector $B = \{B^{ij}\} \in \mathbb{R}^{H_c \times W_c \times C \times 1}$. Finally, we conduct convolution operations on the content feature Z_c with the learned dynamic style kernels F_1, F_2 and B. The output after the convolutions is the feature called $\bar{Z}_{cs}$:

$$\bar{Z}_{cs} = \frac{Z_c - \mu}{\sigma} * F_1 * F_2 + B, \tag{6}$$

where μ and σ are the affine parameters in normalization layer and $*$ is the convolution operator. $\bar{Z}_{cs}$ is then passed to the decoder to generate the stylized image I_{cs}.

Remarks. Note that different from *"style code"* [48] modeled as the parameters (*e.g.* mean and variance) which are shared over all spatial positions globally without taking different semantic regions into consideration, our learned dynamic style kernels are position point-wisely inferred from the globally style-content aligned feature, and therefore are able to fully explore the point-wise semantic structure correlation to modulate local content feature for better artistic style transfer. Our dynamic style kernels make it possible to transfer both the globally aggregated content-style aligned features to local regions in the content images. This ensures our new scheme *"style kernel"* works well on consistent stylization with well preserved structure, and alleviates the shortages of existing style transfer methods [15, 16, 28, 31] which either globally modulate the content feature ignoring local details or overly focus on the local structure details leading to style leakage.

3.3. Loss Functions

The loss function is formulated with an adversarial loss $\mathcal{L}_{adv}$, a reconstruction loss $\mathcal{L}_{rec}$, a REMD loss $\mathcal{L}_{REMD}$, a style loss $\mathcal{L}_{sty}$ [20] and a content loss $\mathcal{L}_{cont}$ [33] as:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cont}\mathcal{L}_{cont} + \lambda_{sty}\mathcal{L}_{sty} + \lambda_{REMD}\mathcal{L}_{REMD} \tag{7}$$

where λ_{adv}, λ_{rec}, λ_{cont}, λ_{sty} and λ_{REMD} are the hyper-parameters used to balance the loss during training.

Reconstruction Loss $\mathcal{L}_{rec}$ is defined as:

$$\begin{aligned}\mathcal{L}_{rec} = \lambda_{rec1}(||I_{cc} - I_c||_2 + ||I_{ss} - I_s||_2) \\ + \lambda_{rec2}\sum_{l=1}^{L}(||\phi_l(I_{cc}) - \phi_l(I_c)||_2 + ||\phi_l(I_{ss}) - \phi_l(I_s)||_2),\end{aligned} \tag{8}$$

where I_{cc} and I_{ss} are the reconstructed images when both the input content and style images are I_c and I_s, respectively. ϕ_l refers to the l_{th} layer in VGG. We chose features from Relu2_1, Relu3_1, Relu4_1 and Relu5_1 layers.

REMD Loss $\mathcal{L}_{REMD}$ [25] adapts the relaxed earth mover distance (REMD) to align the manifold surface of style features. It is formulated as:

$$\mathcal{L}_{REMD} = \max(\frac{1}{W_s H_s}\sum_i \min_j C_{ij}, \frac{1}{W_c H_c}\sum_j \min_i C_{ij}), \tag{9}$$

where C_{ij} denotes the pair-wise cosine distance matrix between the i_{th} and j_{th} feature vector in Z_{cs} and Z_s.

4. Experiments

We train our proposed method on content images from the MS-COCO [30] dataset, and style images from WikiArt [23] database. Each dataset has around 80k images. The training images are first resized to 512, and then 256×256 patches are randomly cropped from the images as inputs. Note that our model can be applied to images of any resolution at the testing stage. Our model is implemented by PyTorch [34]. The Adam [24] is adopted as the optimization solver. We train our model for 160k iterations with a batch size of 16. The learning rate is set to 0.0001.

The hyperparameters in loss functions are set to $\lambda_{cont} = 1$, $\lambda_{rec} = 1$, $\lambda_{sty} = 1$, $\lambda_{rec1} = 20$, $\lambda_{rec2} = 0.5$, $\lambda_{REMD} = 3$, and $\lambda_{adv} = 1$. The encoder is a pretrained VGG-16 network [40], whose parameters are fixed during model training. A multi-scale discriminator is adopted from [43]. We update the discriminator one time after two generator iterations.

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

Qualitative Evaluation. As shown in Figure 4, we present qualitative results of different style transfer methods *i.e.*, AdaConv [2], DRB-GAN [48], statistics matching based method like AdaIN [15], attention based methods including

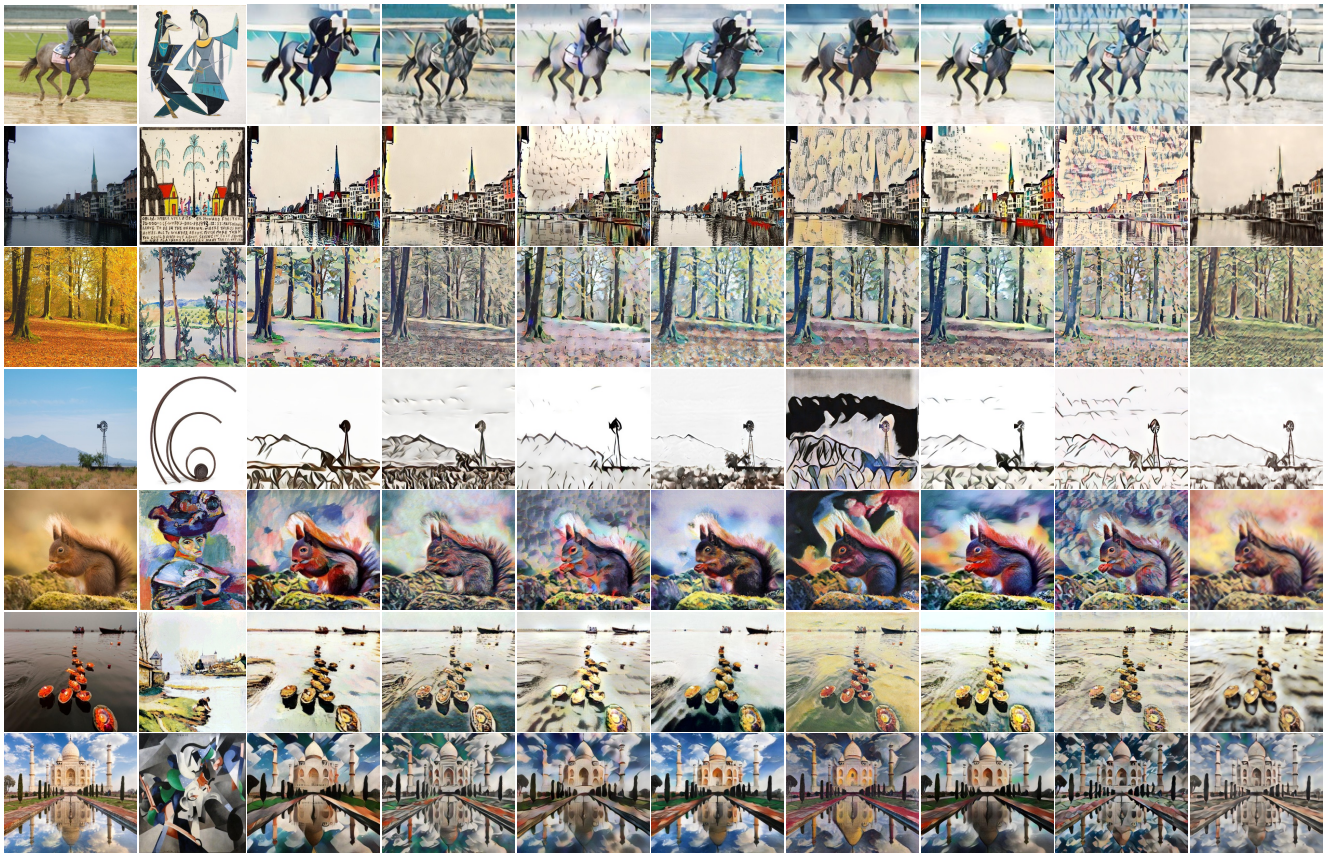

Figure 4. Qualitative performance comparison on stylized results.

Table 2. Quantitative comparison with state-of-the-art methods. We compute the average style loss and the LPIPS score of different methods to indicate how well the style and content are transferred.

Method	AdaIN'17	LST'19	AvatarNet'18	SANet'19	IEC'21	DRB-GAN'21	AdaAttN'21	MAST'21	AdaConv'21	StrTr^2'22	Ours
Sty Loss↓	1.77	2.67	5.91	1.41	1.81	1.51	1.14	-	1.04	1.07	0.98
LPIPS ↓	0.37	0.33	0.34	0.36	0.31	0.33	0.32	0.32	0.42	0.33	0.30

AdaAttN [31] and IEC [5], and feature modification methods like StrTr^2 [8] and MAST [16]. For a fair comparison, we deploy all competing methods for style transfer on images of the smaller dimension resized to 512 while the aspect ratio is preserved. Note that we use the public released code and follow the default configurations for testing.

The comparison in Figure 4 shows the outperformance of our method in terms of visual quality. These images produced by our method faithfully reflect the style characters (*e.g.*, stroke sizes and colors) with no artifact in the regions, and most importantly, they preserve the structural similarity of the content images. On the contrary, AdaIN [15] and MAST [16] fail to generate sharp details and fine strokes. AdaConv [2] yields distorted patterns as it cannot always recover the original style patterns. We also observe non-negligible artificial structures in those images obtained by StrTr^2 [8], DRB-GAN [48] and AdaIN [15]. These methods struggle to preserve the consistency in semantic regions such as the sky. In addition, AdaAttN [31] and IEC [5] and MAST [16] tend to overly preserve the structural similarity and fail to transfer the style to the content image, which makes the results look like the content images.

Quantitative Evaluation. (1) **Style Loss.** Following WCT [28], we adopt the style loss to measure the style consistency between the generated stylizations and the style references. The results of different methods are reported in Table 2. As we can observe, our method achieves the lowest style loss, which indicates that our method mostly transfers the style information to the output. (2) **LPIPS.** We conduct LPIPS to measure the stability and consistency of rendered video clips by following IEC [5]. This metric is to compute the average perceptual distances between consecutive frames and the lower LPIPS score indicates a better stable and consistent performance. We synthesize 10 video clips

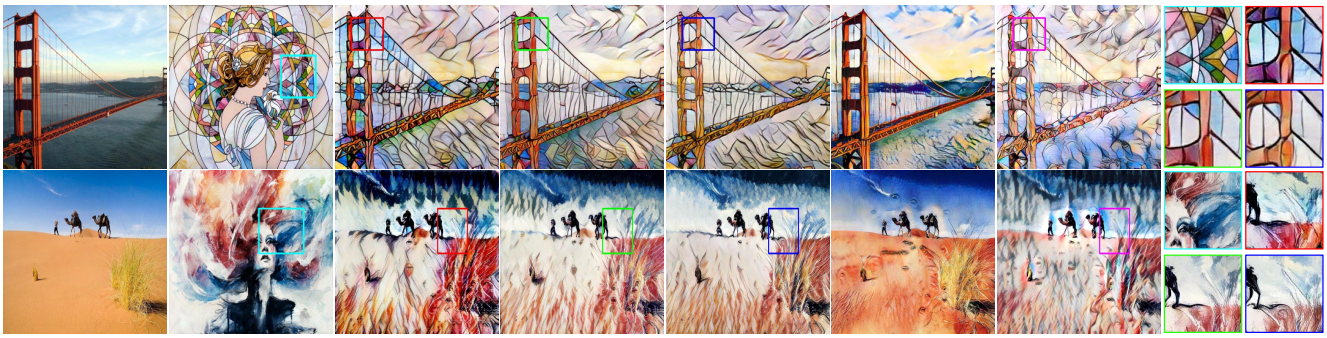

Figure 5. Qualitative performance comparison on stylized results of different model variants. The results of SANet are also listed as reference. *Please zoom in to observe the detailed difference.*

Figure 6. Effectiveness of the Content-based Gating Module (CGM). The results of AdaAttN are also listed as reference.

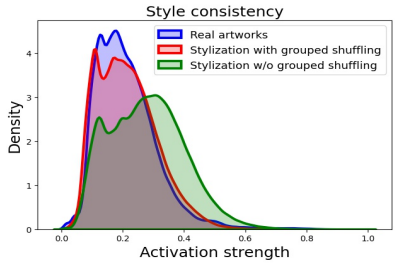

Figure 7. Impact of grouped shuffling: we randomly take 100 content-style pairs to create stylizations. From their feature maps extracted by VGG network, we measure the mean activation value channel-wisely and report the overall distribution.

for each method. In Table 2 we observe that our approach produces the lowest LPIPS score among all methods.

Efficiency Analysis. In Table 1, we compare the inference time of different methods on image resolutions of 256×256 and 512×512. All experiments are conducted using a single Tesla V100 GPU. Our method can achieve 25 fps on 512×512 images, which is comparable with SOTA methods such as AdaAttN and IEC. It worth pointing out that the operation of our style kernel takes FLOPs cost of $H_c \times W_c \times C_{in} \times (k+k+1)$ while vanilla convolution operation has FLOPS cost of $H_c \times W_c \times C_{out} \times (C_{in} \times k \times k+1)$. In Table 3, we compare the inference time given different kernel sizes. Note we set k=3 as a trade-off between the efficiency and performance.

Table 3. Results of different sizes of style kernel.

Setting	Sty Loss↓	LPIPS ↓	T_{256} ↓	T_{512}↓
k=1	1.04	0.30	0.015	0.034
k=3	0.98	0.30	0.017	0.035
k=5	0.97	0.31	0.021	0.044

4.2. Ablation Study

We conduct ablation studies to highlight the effectiveness of different modeling components used by our method.

Effectiveness of Style Kernel. This module works to fully explore the point-wise semantic structure correlation to modulate local content features for better artistic style transfer. As shown in Figure 5, our dynamic style kernels make it possible to transfer both the globally aggregated content- style aligned features to local regions in the content images. This ensures our new scheme "style kernel" works well on consistent stylization with well preserved structure.

Effectiveness of Grouped Shuffling. To verify this module, we report the statistic distribution of activation strength in Table 7. The distribution gap between the results and real artworks is significantly reduced via using grouped shuffling. Without using GS module, the model fails to produce stylizations consistent to the style reference as shown in Figure 5. This demonstrates that the GS shrinks the statistic gap between the generated images and the style images.

Effectiveness of CGM. We summarized the results of our model w/ and w/o CGM in Figure 6. The model w/o CGM produces images containing regions with corrupted patterns and the stylization is not consistent to the style reference. While our model w/ CGM faithfully preserves the completeness of semantic regions. This is because our CGM aims to select focusing regions, where query points can aggregate style information from fairly correlated nodes.

4.3. More Discussions

Generalization of CGM. Our content-based gating modulation can be flexibly deployed as a plug-in module. To ver-

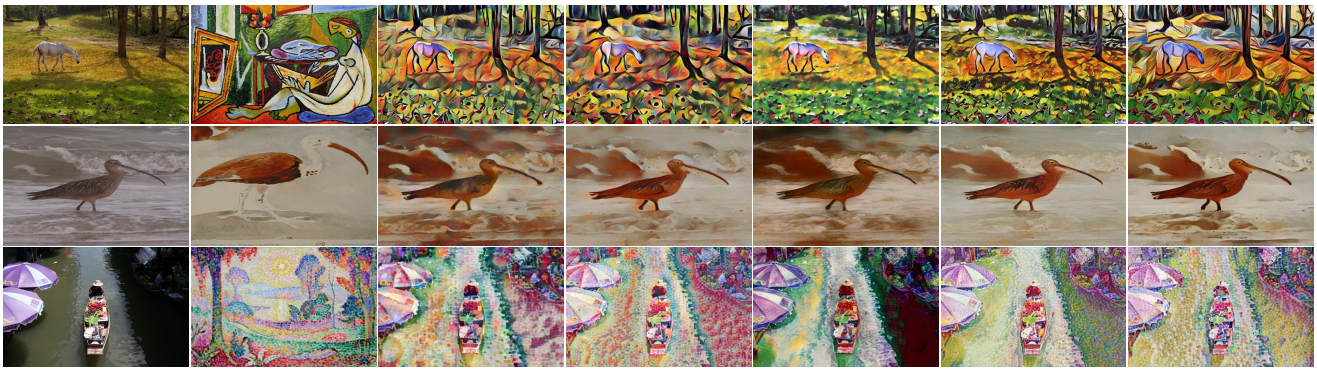

Figure 8. Generalization of our Content-based Gating Module (CGM). By replacing the attention module in SANet and IEC with our CGM, the artifacts and distorted regions in the results of original SANet and IEC are removed and the semantic regions are well preserved.

Figure 9. Comparisons on video style transfer. The 1st column lists 2 frames from a video clip as content images. The style image is at the bottom-right corner.

ify the generalization, we conduct experiments by replacing the attention module in SANet and IEC with our CGM module. The refactored models are denoted as SANet + CGM and IEC + CGM. We train these models with the default settings and compare the performances in Figure 8. We can observe there are nonneglectable artifacts and distorted regions in the results of original SANet and IEC. By plugging-in our CGM, the artifacts are removed and the semantic regions are well preserved. This further proves the effectiveness of our CGM in aggregating semantically consistent information.

Real-world Video Style Transfer. To verify the performance on real-world video stylization, we collect a 1080P video clip consist of 1377 images and compare our method with DRB-GAN [48]. As shown in Figure 9, our approach outperforms existing style transfer methods in terms of style consistency and stability. This can be attributed to the fact that our style kernels are good at transferring the globally aggregated and locally semantic-aligned style features to local regions.

Social Impact and Future Work. Our model achieves impressive image stylizations with well preserved content structure and consistent style characteristics, as shown in Figure 10. This can definitely benefit our society. In the future, we will incorporate text information for text conditioned image stylization.

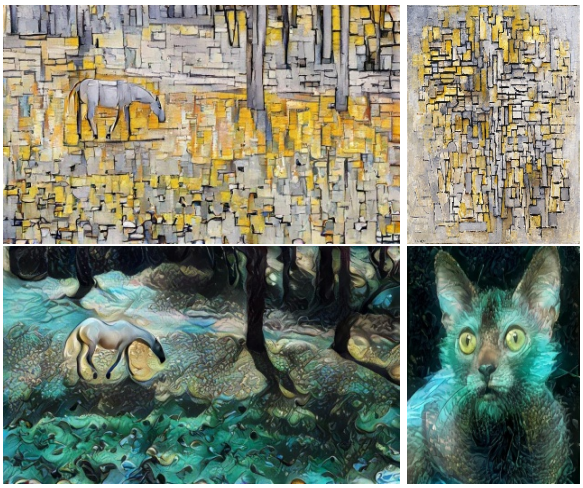

Figure 10. Demonstration of the performance of our method in terms of style consistency and structural similarity.

5. Conclusion

In this paper, we present a new scheme *"style kernel"* for artistic style transfer. Our method learns an attention map with focusing regions using the proposed content-based gating operator. The style feature is then aligned to match the semantics in the content feature based on the learned attention map. In the style kernel generation module, the dynamic parameters *"style kernel"* are learned from the content-style aligned feature and then applied to modulate the content feature for style transfer. Extensive experimental results demonstrate the remarkable performance of our model in generating synthetic style images with better quality than the state-of-the-art.

Acknowledgement

This research was sponsored by Prof. Yongwei Nie's Natural Science Foundation of China (62072191).

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 3

[2] Prashanth Chandran, Gaspard Zoss, Paulo Gotardo, Markus Gross, and Derek Bradley. Adaptive convolutions for structure-aware style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7972–7981, 2021. 2, 3, 5, 6

[3] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 2

[4] 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. *arXiv preprint arXiv:2012.00364*, 2020. 3

[5] 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, 2021. 1, 2, 6

[6] Zhe Chen, Wenhai Wang, Enze Xie, Tong Lu, and Ping Luo. Towards ultra-resolution neural style transfer via thumbnail instance normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 1

[7] Jiaxin Cheng, Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Prem Natarajan. Style-aware normalized loss for improving arbitrary style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 134–143, 2021. 2

[8] 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, 6

[9] Xinzhi Dong, Chengjiang Long, Wenju Xu, and Chunxia Xiao. Dual graph convolutional networks with transformer and curriculum learning for image captioning. In *Proceedings of the ACM International Conference on Multimedia*, 2021. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[11] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven artistic style transfer. In *European Conference on Computer Vision*, pages 717–734. Springer, 2022. 2

[12] 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*, 2016. 2

[13] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020. 3

[14] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. *arXiv preprint arXiv:2102.04378*, 2021. 3

[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*, 2017. 1, 2, 3, 5, 6

[16] Jing Huo, Shiyin Jin, Wenbin Li, Jing Wu, Yu-Kun Lai, Yinghuan Shi, and Yang Gao. Manifold alignment for semantically aligned style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14861–14869, 2021. 1, 2, 5, 6

[17] Yifan Jiang, Bart Wronski, Ben Mildenhall, Jon Barron, Zhangyang Wang, and Tianfan Xue. Fast and high-quality image denoising via malleable convolutions. *arXiv preprint arXiv:2201.00392*, 2022. 3

[18] Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. Dynamic instance normalization for arbitrary style transfer. In *AAAI Conference on Artificial Intelligence*, 2020. 1, 2

[19] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3224–3232, 2018. 3

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[21] Yunjae Jung, Dahun Kim, Sanghyun Woo, Kyungsu Kim, Sungjin Kim, and In So Kweon. Hide-and-tell: Learning to bridge photo streams for visual storytelling. *arXiv preprint arXiv:2002.00774*, 2020. 3

[22] Nikolai Kalischek, Jan D Wegner, and Konrad Schindler. 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*, pages 9382–9391, 2021. 2

[23] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013. 5

[24] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[25] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019. 5

[26] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 1

[27] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast arbitrary style transfer. *arXiv preprint arXiv:1808.04537*, 2018. 2

[28] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in neural information processing systems*, 2017. 2, 5, 6

[29] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022. 3

[30] 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 *European conference on computer vision*, 2014. 5

[31] 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, 5, 6

[32] Songhua Liu, Jingwen Ye, Sucheng Ren, and Xinchao Wang. Dynast: Dynamic sparse transformer for exemplar-guided image generation. In *European Conference on Computer Vision*, pages 72–90. Springer, 2022. 3

[33] 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, 5

[34] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[35] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[36] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021. 3

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[38] Falong Shen, Shuicheng Yan, and Gang Zeng. Neural style transfer via meta networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

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

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 3

[42] Pei Wang, Yijun Li, and Nuno Vasconcelos. Rethinking and improving the robustness of image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 124–133, 2021. 2

[43] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 5

[44] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *IEEE ICCV*, 2021. 3

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3

[46] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 3

[47] 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

[48] Wenju Xu, Chengjiang Long, Ruisheng Wang, and Guanghui Wang. Drb-gan: A dynamic resblock generative adversarial network for artistic style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6383–6392, 2021. 1, 2, 3, 5, 6, 8

[49] Jiaqi Yu, Yongwei Nie, Chengjiang Long, Wenju Xu, Qing Zhang, and Guiqing Li. Monte carlo denoising via auxiliary feature guided self-attention. *ACM Transactions on Graphics*, 40(6), 2021. 3

[50] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *arXiv preprint arXiv:2111.09881*, 2021. 3

[51] 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

[52] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2482–2491, 2019. 3