AdaAttN: Revisit Attention Mechanism in Arbitrary Neural Style Transfer

Songhua Liu[1,2,*], Tianwei Lin[1,†], Dongliang He[1], Fu Li[1], Meiling Wang[1],
Xin Li[1], Zhengxing Sun[2,†], Qian Li[3], Errui Ding[1]
[1]Department of Computer Vision Technology (VIS), Baidu Inc.,
[2]Nanjing University, [3]National University of Defense Technology
[1]{liusonghua,lintianwei01,hedongliang01,lifu,wangmeiling03,lixin41,dingerrui}@baidu.com,
[2]songhua.liu@smail.nju.edu.cn, szx@nju.edu.cn, [3]liqian@nudt.edu.cn

Abstract

Fast arbitrary neural style transfer has attracted widespread attention from academic, industrial and art communities due to its flexibility in enabling various applications. Existing solutions either attentively fuse deep style feature into deep content feature without considering feature distributions, or adaptively normalize deep content feature according to the style such that their global statistics are matched. Although effective, leaving shallow feature unexplored and without locally considering feature statistics, they are prone to unnatural output with unpleasing local distortions. To alleviate this problem, in this paper, we propose a novel attention and normalization module, named ***Ada****ptive* ***Att****ention* ***N****ormalization (****AdaAttN****), to adaptively perform attentive normalization on per-point basis. Specifically, spatial attention score is learnt from both shallow and deep features of content and style images. Then per-point weighted statistics are calculated by regarding a style feature point as a distribution of attention-weighted output of all style feature points. Finally, the content feature is normalized so that they demonstrate the same local feature statistics as the calculated per-point weighted style feature statistics. Besides, a novel local feature loss is derived based on AdaAttN to enhance local visual quality. We also extend AdaAttN to be ready for video style transfer with slight modifications. Experiments demonstrate that our method achieves state-of-the-art arbitrary image/video style transfer. Codes and models are available on https://github.com/wzmsltw/AdaAttN.*

1. Introduction

Given a content image I_c and a style image I_s, artistic style transfer aims at applying style patterns of I_s onto I_c while preserving content structure of I_c simultaneously, which is widely used in computer-aid art generation. The seminal work of Gatys *et al*. [9] proposed an image optimization method that iteratively minimizes the joint content and style loss in the feature space of a pre-trained deep neural network. This time-consuming optimization process has motivated researchers to explore more efficient approaches. Johnson *et al*. [18] alternatively considered using a feed-forward network to generate rendered images directly and enabled real time style transfer. Since the learned model can only work for one specific style, this method and its following works [40, 34, 35, 21, 27, 39, 16, 20] are categorized to *Per-Style-Per-Model* method [17]. In the literature, there are *Multiple-Style-Per-Model* solutions [7, 2, 23, 44] and *Arbitrary-Style-Per-Model* [14, 3, 22, 28, 15, 24, 6, 5, 43, 32, 42, 11] methods. In the latter case, a model can accept any style image as input and produce stylized results in a single forward pass once upon the model is trained. Therefore, it is the most flexible and attracts increasing attention from academic, industrial, and art communities.

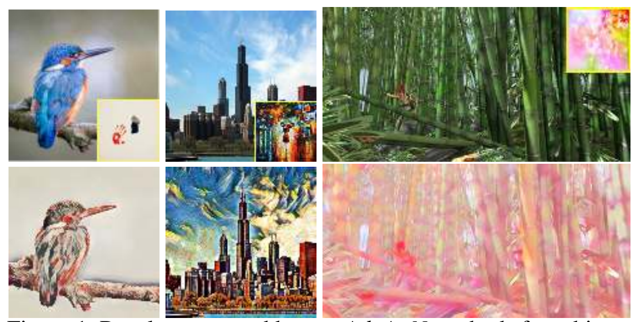

Figure 1. Results generated by our *AdaAttN* methods for arbitrary image/video style. Animated clips at the right side can be found in our supplementary material.

Nevertheless, arbitrary style transfer is far from being solved. Enabling flexibility sacrifices local style pattern modeling capability for an arbitrary style transfer network. For example, the pioneering work [14] proposed a simple yet effective method *AdaIN*, which transfers global mean and variance of a style image to a content image in the

*This work was done when Songhua Liu was an intern at VIS, Baidu.
†Corresponding authors.

feature space to support arbitrary input style image. Since mean and variance of features are calculated globally, local details and point-wise patterns are largely dismissed and thus the local stylization performance is largely degraded [28]. Similar trade-off between flexibility and capability also exists in [5, 22, 15, 24, 10], where all local feature points of the content image are processed by the same transformation function based on style images. To enhance the locality awareness of arbitrary style transfer models, recently, attention mechanism is adopted in multiple works [28, 6, 43] for this task. Their common intuition is that a model should pay more attention to those feature-similar areas in the style image for stylizing a content image region. Such attention mechanism has been proved to be effective for generating more local style details in arbitrary style transfer. Unfortunately, while improving the performance, it fails to totally solve this problem and the local distortions still occur.

It is not so hard to reveal the reasons of the above dilemma posed to the attention mechanism. Digging into the details of current attention based solutions for arbitrary style transfer, it can be easily figured out that 1) the designed attention mechanisms are commonly based on deep CNN features on higher abstraction levels and the low-level details are dismissed; 2) the attention scores are usually used to re-weight feature maps of the style image and the re-weighted style feature is simply fused into content feature for decoding. Deep CNN features based attention strategy leaves the low-level patterns of images at shallow network layers unexplored. Thus the attention scores may focus little on low-level textures and are dominated by high-level semantics. Meanwhile, the spatial re-weighting of style features followed by fusion of re-weighted style features and content features, as done in SANet [28] (Figure 3(b)), works without consideration of feature distribution.

To this end, we attempt to address these issues and get a better balance between style pattern transferring and content structure preserving. Motivated by lessons learned through the above analysis, we propose a novel attention and normalization module named ***Ada****ptive* ***Att****ention* ***N****ormalization (**AdaAttN**)* for arbitrary style transfer. It can adaptively perform attentive normalization on per-point basis for feature distribution alignment. In more detail, spatial attention score is learnt from both shallow and deep features of content and style images. Then, per-point weighted statistics is calculated by regarding a style feature point as a distribution of attention-weighted output of all spatial feature points. Finally the content feature is normalized so that its local feature statistics are the same as the per-point weighted style feature statistics. In this way, the attention module takes into account both shallow and deep CNN features of the style image as well as the content image. Meanwhile, alignment of per-point feature statistics from the content feature to the modulated style feature is achieved. Based on *AdaAttN* module, a novel optimization objective named *local feature loss* and a new arbitrary image style transfer pipeline are derived. Our contributions can be summarized as follows:

- We introduce a novel *AdaAttN* module for arbitrary style transfer. It takes both shallow and deep features into account for attention score calculation and properly normalizes content feature such that feature statistics are well aligned with attention-weighted mean and variance maps of style features on per-point basis.
- A new optimization objective called *local feature loss* is proposed. It helps the model training and improves arbitrary style transfer quality by regularizing local features of the generated image.
- Extensive experiments and comparisons with other state-of-the-art methods are performed to demonstrate the effectiveness of our proposed method.
- Further extension of our model for video style transfer via simply introducing cosine-distance based attention and image-wise similarity loss can result in stable and appealing results.

2. Related Works

2.1. Arbitrary Style Transfer

Recent arbitrary style transfer methods can be divided into two categories: global transformation based and local transformation based methods. The common idea of the former category is to apply feature modification globally. WCT [24] achieved this with two transformation steps including whitening and coloring. Huang *et al*. [14] proposed AdaIN that adaptively applies mean and standard deviation of each style feature to shift and re-scale the corresponding normalized content feature so that content feature and style feature share the same distribution. Jing *et al*. [15] extended this method by dynamic instance normalization, where weights for intermediate convolution blocks are generated by another network taking the style image as input. Li *et al*. [22] proposed to generate a linear transformation according to content and style features. Furthermore, Deng *et al*. [5] got the transformation function with multi-channel correlation. Although these methods accomplish the overall arbitrary style transfer task and make great progress in this field, local style transfer performance is generally unsatisfactory due to global transformations leveraged by them are hard to take care of detailed local information.

For the latter, Chen *et al*. [3] proposed a style-swap method, which is a patch based style transfer method relying on similarities between content and style patches. [11] was another patch based method considering matching of both global statistics and local patches. Avatar-Net [32] further proposed a multi-scale framework that combines ideas of style swap and AdaIN function. In recent years, attention mechanism is widely used in arbitrary style transfer thanks to its excellent ability to model fine grained correspondence among local features of style and content images. On this routine, Park *et al*. [28] proposed Style-Attentional Network (*SANet*) to match content and style features. Yao *et al*. [43] considered different types of strokes with such attention framework. Deng *et al*. [6] proposed a multi-adaptation module that applies point-wise attention for content features and channel-wise attention for style features. Common practices adopted by these methods are to build the attention mechanism merely upon deep CNN features without considering shallow features and simply mix the content feature and the re-weighted style feature. Thus, it tends to distort original content structures largely and results in undesired effect for human eyes. In this paper, our goal is to explore a better trade-off between style pattern transferring and content structure preserving.

2.2. Video Style Transfer

Directly applying image style transfer techniques on video frame sequences usually results in flickering effects caused by temporal inconsistency. Thus, a lot of works add optical flow consistency constraint to original image style transfer solutions, *e.g*., [30] for optimization based video style transfer, [31, 1, 12, 13, 8] for per-style-per-model methods, [37, 38] for arbitrary-style-per-model methods, and [36, 4, 26] for image-to-image translation frameworks. Optical flow constraint improves the stability of video style transfer. However, it heavily relies on a pre-extracted optical flow field with high accuracy to perform flow-based warping. There are also some works that address the stability issue with approaches other than optical flow warping. [22, 5] leveraged linearity of transformation models to guarantee inter-frame consistency on feature space. Wu *et al*. [41] proposed a *SANet* based method that leads current frame to focus on similar regions of previous frame with the help of a *SSIM* consistency constraint. Different from these methods, in this work, we add a novel image-wise similarity loss based on attention mechanism to overcome the flickering artifact and comparable or even better stability is achieved without prerequisite optical flow.

3. Methods

3.1. Overall Framework

The proposed network takes a style image I_s and a content image I_c to synthesize a stylized image I_{cs}. In our proposed model, we employ a pre-trained VGG-19 network [33] as encoder to extract multi-scale feature maps. The decoder follows the setting of [14] with a symmetric structure of VGG-19. In order to take full advantage of features in both shallow and deep levels, we employ a multi-level strategy by integrating three AdaAttN modules on *ReLU-3_1* , *ReLU-4_1* and *ReLU-5_1* layers of VGG, respectively, as shown in Figure 2. We denote the extracted feature of layer *ReLU-x_1* in VGG as $F_*^x \in R^{C \times H_* W_*}$ when it takes an image I_* as input and $*$ can be c or s here representing content and style features respectively. To fully exploit low-level patterns, we further concatenate feature of current layer with down-sampled features of its previous layers as:

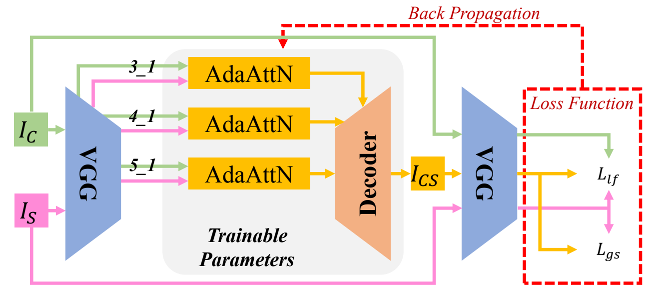

Figure 2. Overview of full framework, where three *AdaAttN* modules and the decoder are trainable. $\mathcal{L}_{lf}$ and $\mathcal{L}_{gs}$ are local feature loss and global style loss separately.

$$F_*^{1:x} = D_x(F_*^1) \oplus D_x(F_*^2) \oplus \cdots \oplus F_*^x, \tag{1}$$

where D_x stands for the bilinear interpolation layer which downsamples the input feature to the same shape of F_*^x, and $\oplus$ here means concatenation operation along channel dimension. Then, we can denote the embedded feature of the AdaAttN module at layer l as:

$$F_{cs}^x = AdaAttN(F_c^x, F_s^x, F_c^{1:x}, F_s^{1:x}), \tag{2}$$

where F_c, F_s and F_{cs} are content, style, and embedded feature, respectively. With multi-level embedded features, we can synthesize the stylized image I_{cs} with decoder as:

$$I_{cs} = Dec(F_{cs}^3, F_{cs}^4, F_{cs}^5). \tag{3}$$

3.2. Adaptive Attention Normalization

The feature transformation module is the key component in arbitrary style transfer models. A comparison of our module with other frameworks is demonstrated in Figure 3. The pioneering AdaIN [14] only considers the holistic style distributions and the content feature is manipulated such that its feature distribution globally aligns with that of the style feature. By taking local style patterns into consideration, SANet [28] calculates attention map from the style and content feature maps and then modulates the style feature with the attention map to fuse the attention output into the content feature. SANet performs in local stylization. However, it lacks low-level matching and local feature distribution alignment. Inspired by lessons learned from AdaIN

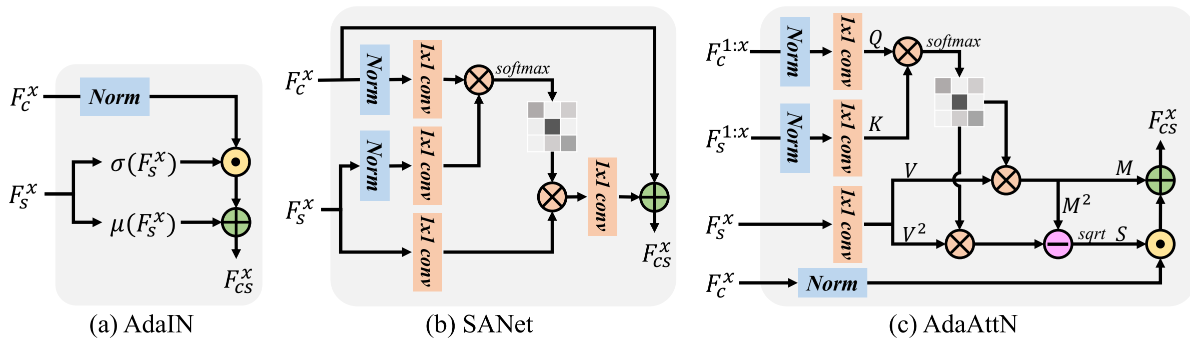

Figure 3. (a) The structure of *AdaIN* [14] module; (b) The structure of *SANet* [28] module; (c) The structure of our proposed *AdaAttN* module. $Norm$ here denotes the mean-variance channel-wise normalization.

and SANet, we propose the *Adaptive Attention Normalization (AdaAttN)* module, which can adaptively transfer feature distribution on per-point basis via considering both low-level and high-level features with attention mechanism. As shown in Figure 3(c), AdaAttN works in three steps: **(1)** computing attention map with content and style features from shallow to deep layers; **(2)** calculating weighted mean and standard variance maps of style feature; **(3)** adaptively normalizing content feature for per-point feature distribution alignment.

Attention Map Generation. In arbitrary style transfer methods, attention mechanism is used to measure the similarity between content and style features. Different from previous methods which only use relatively deep features, we use low-level and high-level layers of both content and style feature simultaneously. To computing attention map A of layer x, we formulate Q (query), K (key) and V (value) as:

$$\begin{aligned} Q &= f(Norm(F_c^{1:x})), \\ K &= g(Norm(F_s^{1:x})), \\ V &= h(F_s^x), \end{aligned} \tag{4}$$

where f, g, and h are 1×1 learnable convolution layers, $Norm$ here denotes channel-wise mean-variance normalization, as used in instance normalization. The attention map A can be calculated as:

$$A = Softmax(Q^\top \otimes K), \tag{5}$$

where $\otimes$ denotes matrix multiplication.

Weighted Mean and Standard Variance Map. Applying attention score matrix A to style feature F_s^x as SANet [28] does can be regarded as calculating each target style feature point by weighted summation of all style feature points. In this paper, we interpret this process as viewing a target style feature point by attention output as a distribution of all the weighted style feature points. Then from this perspective, we can calculate statistics for each distribution. We term such statistics as *attention-weighted mean* and *attention-weighted standard variance* respectively. Thus, the attention-weighted mean $M \in R^{C \times H_c W_c}$ becomes:

$$M = V \otimes A^\top, \tag{6}$$

where $A \in R^{H_c W_c \times H_s W_s}$ and $V \in R^{C \times H_s W_s}$. Since variance of a variable equals to the expectation of its square minus the square of its expectation, we can obtain the attention-weighted standard deviation $S \in R^{C \times H_c W_c}$ as:

$$S = \sqrt{(V \cdot V) \otimes A^\top - M \cdot M}, \tag{7}$$

where $\cdot$ denotes element-wise product.

Adaptive Normalization. Finally, for each position and each channel of normalized content feature map, corresponding scale in S and shift in M are used to generate transformed feature map:

$$F_{cs}^x = S \cdot Norm(F_c^x) + M. \tag{8}$$

In short, AdaAttN performs feature statistics transferring via generating attention-weighted mean and variance maps. As shown in Figure 3, compared with AdaIN, AdaAttN considers per-point statistics rather than globally. AdaAttN is more general than AdaIN. For each i, j, if set $A_{i,j} = 1/(H_s W_s)$, AdaAttN can be specialized to AdaIN. Compared with SANet, attention mechanism is used to calculate target feature distribution instead of directly generating transferred feature for further fusion.

3.3. Loss Function

Our overall loss function is the weighted summation of global style loss ($\mathcal{L}_{gs}$) and local feature loss ($\mathcal{L}_{lf}$):

$$\mathcal{L} = \lambda_g \mathcal{L}_{gs} + \lambda_l \mathcal{L}_{lf}, \tag{9}$$

where λ_g and λ_l are hyper-parameters controlling weights of their corresponding loss terms. Details of each loss term will be explained in the remaining part of this section.

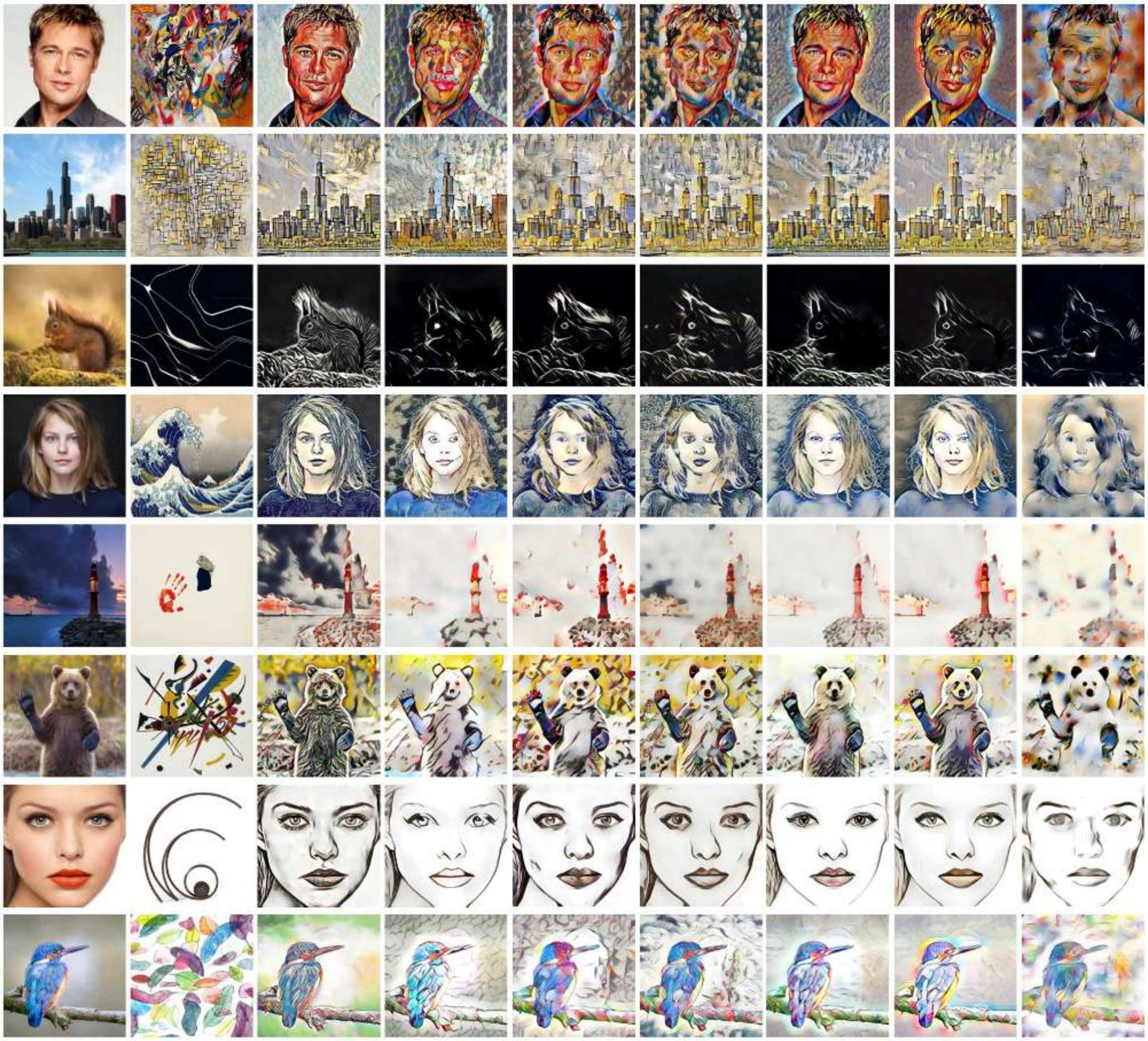

Figure 4. Comparison with other state-of-the-art methods in arbitrary image style transfer.

To begin with, following [14] and many other works, distances of mean μ and standard deviation σ between generated image and style image in VGG feature space are penalized to guarantee global stylized effect ($\mathcal{L}_{gs}$):

$$\mathcal{L}_{gs} = \sum_{x=2}^{5}(||\mu(E^x(I_{cs})) - \mu(F_s^x)||_2 + ||\sigma(E^x(I_{cs})) - \sigma(F_s^x)||_2), \tag{10}$$

where $E()$ denotes feature of the VGG encoder and its superscript ^{x} denotes the layer index.

The proposed novel local feature loss $\mathcal{L}_{lf}$ constrains that feature map of stylized image is consistent with the transformation result by *AdaAttN* function:

$$\mathcal{L}_{lf} = \sum_{x=3}^{5}\left|\left|E^x(I_{cs}) - AdaAttN^*(F_c^x, F_s^x, F_c^{1:x}, F_s^{1:x})\right|\right|_2, \tag{11}$$

where $AdaAttN^*$ serves as a supervision signal that should be deterministic. Thus, we consider the parameter-free version of $AdaAttN$ without the three learnable 1×1 convolution kernels (f, g, and h). Local feature loss makes the model generates better stylized output for local areas compared with conventional content loss term used in [14, 28].

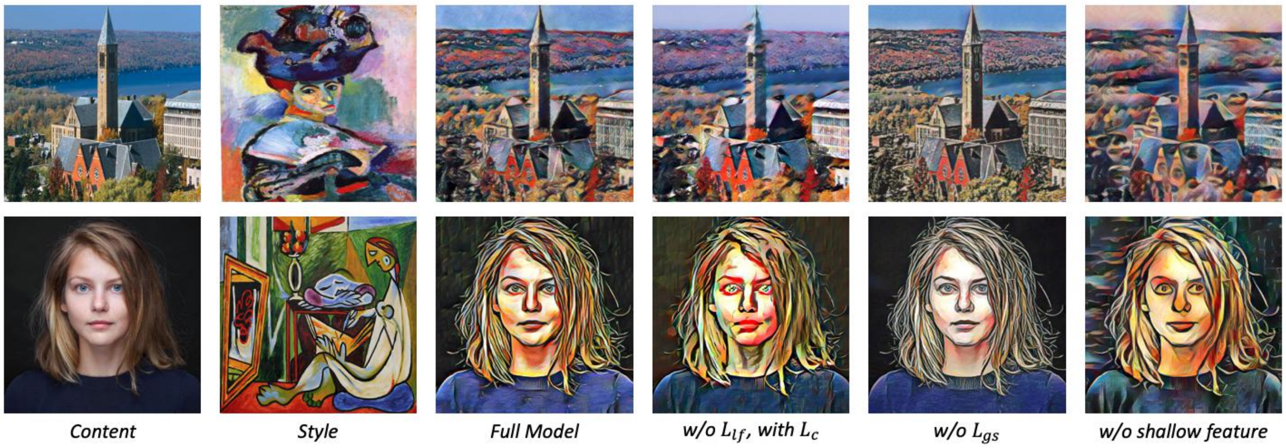

Figure 5. Ablation study on loss functions and shallow feature. Zoom-in for better view.

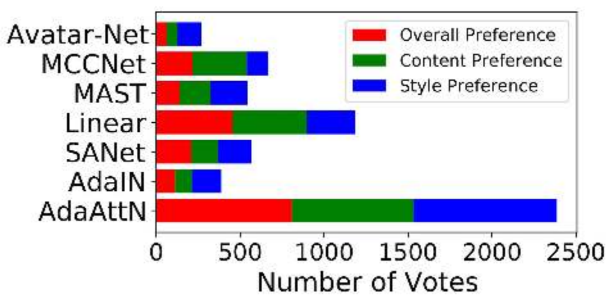

Figure 6. Results of user study.

Method	Inference Time (sec./image)		
	256×256	512×256	512×512
Avatar-Net	0.124	0.176	0.311
AdaIN	0.038	0.049	0.066
Linear	0.028	0.036	0.049
MCCNet	0.024	0.040	0.057
MAST	0.046	0.073	0.115
SANet	0.043	0.064	0.081
Ours	0.051	0.066	0.112

Table 1. Running speed comparison under different resolutions.

3.4. Extension for Video Style Transfer

Compared to other attention-based methods, our method is capable of generating more natural stylized results without much local distortions, and thus it is of great potential for video style transfer. With two slight modifications, our model can be extended to video style transfer.

First of all, we notice that *Softmax* function in Eq.5 shows strong exclusiveness in attention score due to exponential computation and it can focus majorly on local patterns and has negative influence in stabilization. For video style transfer, alternatively, we consider cosine similarity for attention map computation:

$$A_{i,j} = \frac{S_{i,j}}{\sum_j S_{i,j}}, S_{i,j} = \frac{Q_i \cdot K_j}{||Q_i|| \times ||K_j||} + 1, \tag{12}$$

where cosine similarity results in more flat attention score distribution than Softmax. Thus, the local feature statistics will be more stable and local style patterns will not be over emphasised, leading to better assurance of consistency.

Secondly, based on attention mechanism, we design a novel cross-image similarity loss term $\mathcal{L}_{is}$ to regularize the relevant contents between two content images c_1, c_2:

$$\mathcal{L}_{is} = \sum_{x=2}^{4} \frac{1}{N_{c_1}^x N_{c_2}^x} \sum_{i,j} \left| \frac{D_{c_1,c_2}^{i,j,x}}{\sum_i D_{c_1,c_2}^{i,j,x}} - \frac{D_{cs_1,cs_2}^{i,j,x}}{\sum_i D_{cs_1,cs_2}^{i,j,x}} \right|,$$
$$D_{u,v}^{i,j,x} = 1 - \frac{F_u^{x,i} \cdot F_v^{x,j}}{\left|\left|F_u^{x,i}\right|\right| \times \left|\left|F_v^{x,j}\right|\right|}, \tag{13}$$

where N_c^x is the size of spatial dimension in content feature map F_c^x of layer *ReLU-x_1*, $F_*^{x,i}$ means feature vector of i-th position of F_*^x, and $D_{u,v}^{i,j,x}$ measures cosine distance of $F_u^{x,i}$ and $F_v^{x,j}$. In each training iteration, two input video frames are sampled to enable this loss. Intuitively, such cross-image similarity loss requires the stylized results of two content images share similar local similarity patterns with the two original images. Therefore, it ensures awareness of inter-frame relationship in video style transfer and contributes to stable results.

4. Experiments

4.1. Implementing Details

We train our arbitrary style transfer model with *MS-COCO* [25] as our content image set and *WikiArt* [29] as our style image set. λ_g, λ_l, and λ_{is} (for video style transfer

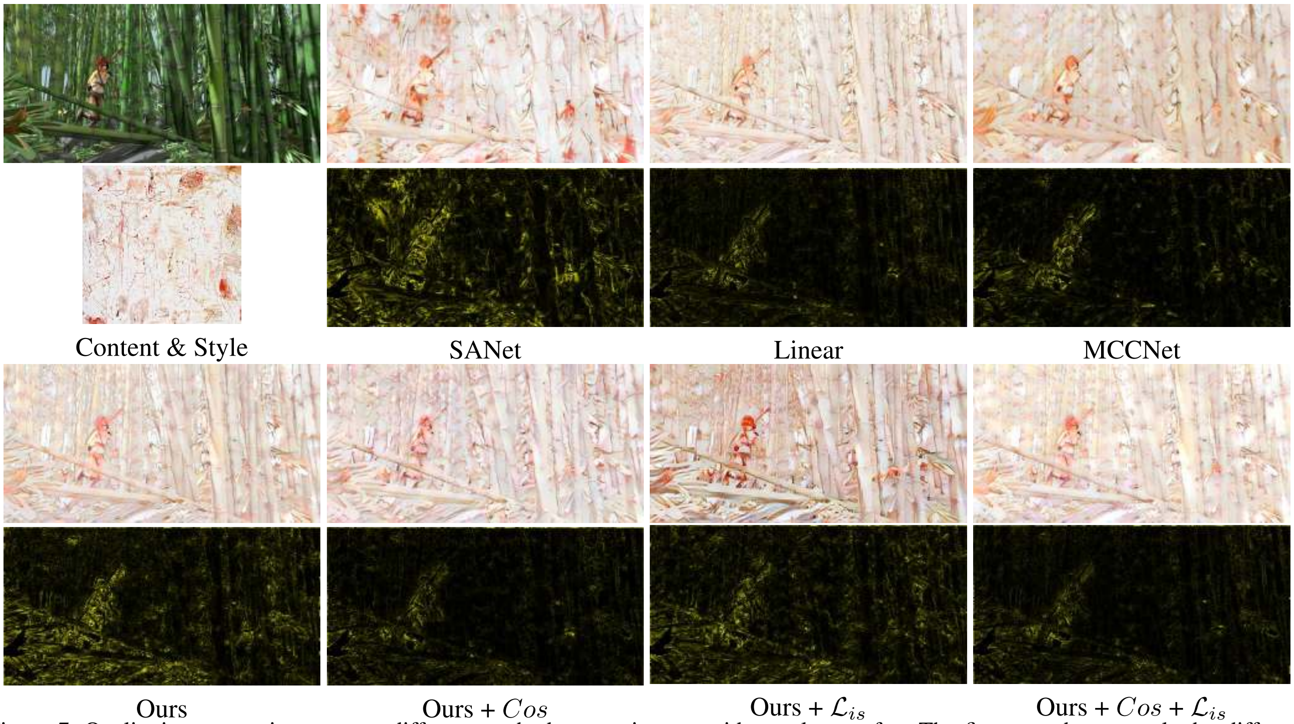

Figure 7. Qualitative comparison among different methods or settings on video style transfer. The first row shows results by different methods or settings. The second row shows the corresponding optical flow error map.

only) are set as 10, 3, and 100, respectively. *Adam* [19] with α, β_1, and β_2 of 0.0001, 0.9, and 0.999, is used as solver. In the training phase, all images are loaded with 512×512 resolution and randomly cropped to 256×256 for augmentation. While inference, our model can be applied to images with any resolution. In this section, 512×512 and 512×256 resolutions are used for image and video, respectively. The training lasts for 50K iterations on a single Nvidia Tesla P40 GPU and batch size is 8 for image and 4 for video. Please refer to supplementary for detailed network configurations.

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

Qualitative Comparison. As shown in Figure 4, we compare our method with six state-of-the-art arbitrary style transfer methods, including AdaIN [14], SANet [28], MAST [5], Linear [22], MCCNet [5], and Avatar-Net [32]. AdaIN [14] directly adjusts second-order statistics of content feature globally and we can see style patterns are transferred with severe content details lost (1st, 5th and 6th rows). Avatar-Net [32] utilizes AdaIN for multi-scale transferring and introduces style decorators with patch matching strategies, which results in blur stylization results with obvious plaques (1st, 6th and 8th rows). SANet [28] and MAST [5] adopt attention mechanism to attentively transfer style features to content features in deep layers. It will result in damaged content structures (3rd, 4th and 6th rows) and dirty textures (1st, 2nd, 8th rows). Some style patches are even directly transferred into the content image improperly (4th, 8th rows). Linear [22] and MCCNet [5] modify features via linear projection and per-channel correlation respectively, both resulting in relative clean stylization outputs. However, textural patterns of style images are not captured adaptively, loss of content details is encountered (3rd, 5th and 6th rows), and content image color remains (7th row). As shown in the 3rd column, AdaAttN can adaptively transfer style patterns to each location of content images appropriately, attributing to the novel attentive normalization on per-point. It shows that AdaAttN achieves a better balance between style transferring and content structure preserving.

User Study. Following SANet, 15 content images and 15 style images are randomly picked to form 225 images pairs in total. Then we randomly sample 20 content-style pairs and synthesize stylized images by different methods. Results are presented side-by-side in a random order and we ask subjects to select their favorite one from three views: content preservation, stylization degree, and overall preference. We collect 2000 votes for each view from 100 users and show the number of votes of each method in the form of bar chart. The results in Figure 6 demonstrate that our stylized results are more appealing than competitors.

Efficiency Analysis. We demonstrate run time performance of AdaAttN and SOTA feed-forward methods on Table 1. All experiments are conducted using a single Nvidia P40 GPU. Although multi-depth feature layers (from 1_1 to 5_1) are used, our method can still achieve 20 FPS at 256px, which is comparable with SOTA methods such as SANet [28] and Linear [22]. Thus, our proposed AdaAttN can practicably synthesize stylized images in real time.

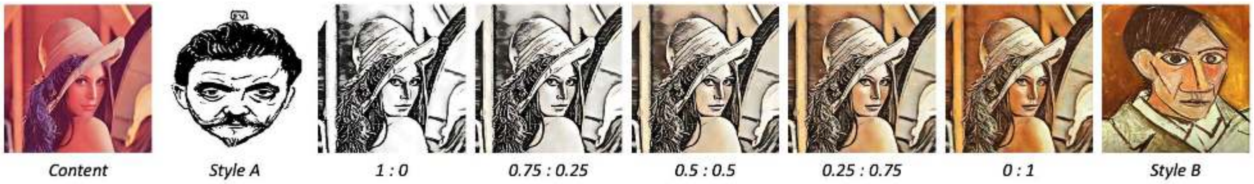

Figure 8. Style interpolation.

Method	Style1	Style2	Style3	Style4	Mean
SANet	8.57	8.93	10.3	4.66	7.76
Linear	4.41	5.10	5.24	2.67	4.42
MCCNet	4.63	4.84	5.48	2.35	4.45
Ours	5.65	5.77	6.41	3.39	5.52
Ours + Cos	4.09	4.59	5.15	2.26	4.09
Ours + $\mathcal{L}_{is}$	5.51	5.31	6.26	3.31	5.51
Ours + Cos + $\mathcal{L}_{is}$	**3.70**	**4.46**	**4.49**	**2.14**	**3.91**

Table 2. Optical flow error ($\times 10^{-2}$) of SOTA methods and different AdaAttN variants. Smaller values mean better temporal consistency. Cos here stands for attention score of cosine similarity. The mean value is calculated using 20 styles.

4.3. Ablation Study

Loss Function. As shown in Figure 5, we present ablation study results to verify the effectiveness of each loss term used for training AdaAttN. **(1)** To verify the effectiveness of our proposed local feature loss $\mathcal{L}_{lf}$, we replace it with the vanilla *L2* content loss $\mathcal{L}_c$ which constraints feature distance between I_c and I_{cs} and is used in many style transfer methods [14, 6, 28]. As shown in the 4th column, their visual quality is obviously worse than that of the full model. It suggests that compared to content loss, our proposed local feature loss can better take style patterns into consideration while preserving content structures. **(2)** We remove the global style loss $\mathcal{L}_{gs}$ and train model with $\mathcal{L}_{lf}$ only. As shown in the 5th column, style patterns are also weakly transferred without style loss, which demonstrates that $\mathcal{L}_{lf}$ can force network to learn style transfer to some extent. However, the overall color saturation is degraded, showing that the global style loss is necessary.

Low-level Feature. To verify the effectiveness of shallow feature used in AdaAttN, we remove shallow feature via replacing the Q and K of AdaAttN from $F^{1:x}$ to F^x. Some local content damage and dirty textures can be observed (last column of Figure 5). Our AdaAttN can effectively utilize shallow features to generate pleasant stylization results.

4.4. Video Style Transfer

For video stylization, we compare our method with SOTA methods SANet, Linear and MCCNet, where optical flow is not used for stabilization. To verify the effectiveness of our proposed methods for video stylization, we also provide ablation results of adding Cos and $\mathcal{L}_{is}$, where Cos denotes attention score of cosine similarity (Eq.12). The qualitative results in Figure 7 and quantitative results in Table. 2 both demonstrate that (1) our method is more stable than attention-based method *SANet*; (2) replacing Softmax activation with cosine distance based attention can significantly improve temporal consistency; (3) with our proposed modifications, *AdaAttN* is more stable than *Linear* and *MCCNet*, which are proposed for video stylization.

Figure 9. Results of multi-style transfer.

4.5. Multi-Style Transfer

Following previous works [28, 6], we explore interpolating several style images via averaging their mean and standard variance maps of different styles, then the combined mean and variance are used to modulate the content feature for decoding (Figure 8). Besides style interpolation, we can also achieve multi-style transfer via concatenating multiple style images into one image and feeding it into AdaAttN (Figure 9). From these results, we can see AdaAttN can flexibly support various run-time controls with plausible outcomes.

5. Conclusion

In this paper, we propose a novel *AdaAttN* module for arbitrary style transfer. AdaAttN performs feature statistics transferring via modulation with per-point attention-weighted mean and variance of style feature. Attention weights are generated from both style and content features from low-level to high-level. With slight modifications, our model is ready for video style transfer as well. Experiment results demonstrate that our method can generate high quality stylization results for both images and videos. AdaAttN has the potential for improving other image manipulation or translation tasks, we will explore this in our future work.

References

[1] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. Coherent online video style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1105–1114, 2017. 3

[2] 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*, pages 1897–1906, 2017. 1

[3] Tian Qi Chen and Mark Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*, 2016. 1, 2

[4] Yang Chen, Yingwei Pan, Ting Yao, Xinmei Tian, and Tao Mei. Mocycle-gan: Unpaired video-to-video translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 647–655, 2019. 3

[5] Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. Arbitrary video style transfer via multi-channel correlation. *arXiv preprint arXiv:2009.08003*, 2020. 1, 2, 3, 7

[6] Yingying Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. Arbitrary style transfer via multi-adaptation network. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2719–2727, 2020. 1, 2, 3, 8

[7] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 1

[8] Chang Gao, Derun Gu, Fangjun Zhang, and Yizhou Yu. Reconet: Real-time coherent video style transfer network. In *Asian Conference on Computer Vision*, pages 637–653. Springer, 2018. 3

[9] 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 (CVPR)*, June 2016. 1

[10] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv preprint arXiv:1705.06830*, 2017. 2

[11] Shuyang Gu, Congliang Chen, Jing Liao, and Lu Yuan. Arbitrary style transfer with deep feature reshuffle. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8222–8231, 2018. 1, 2

[12] Agrim Gupta, Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Characterizing and improving stability in neural style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4067–4076, 2017. 3

[13] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. Real-time neural style transfer for videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 783–791, 2017. 3

[14] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 1, 2, 3, 4, 5, 7, 8

[15] Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. Dynamic instance normalization for arbitrary style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4369–4376, 2020. 1, 2

[16] Yongcheng Jing, Yang Liu, Yezhou Yang, Zunlei Feng, Yizhou Yu, Dacheng Tao, and Mingli Song. Stroke controllable fast style transfer with adaptive receptive fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 238–254, 2018. 1

[17] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385, 2019. 1

[18] 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. 1

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[20] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[21] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European conference on computer vision*, pages 702–716. Springer, 2016. 1

[22] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast arbitrary style transfer. *arXiv preprint arXiv:1808.04537*, 2018. 1, 2, 3, 7

[23] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Diversified texture synthesis with feed-forward networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3920–3928, 2017. 1

[24] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *arXiv preprint arXiv:1705.08086*, 2017. 1, 2

[25] 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*, pages 740–755. Springer, 2014. 6

[26] Songhua Liu, Hao Wu, Shoutong Luo, and Zhengxing Sun. Stable video style transfer based on partial convolution with depth-aware supervision. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2445–2453, 2020. 3

[27] Xiao-Chang Liu, Ming-Ming Cheng, Yu-Kun Lai, and Paul L Rosin. Depth-aware neural style transfer. In *Proceedings of the Symposium on Non-Photorealistic Animation and Rendering*, pages 1–10, 2017. 1

[28] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5880–5888, 2019. 1, 2, 3, 4, 5, 7, 8

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

[30] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. In *German conference on pattern recognition*, pages 26–36. Springer, 2016. 3

[31] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos and spherical images. *International Journal of Computer Vision*, 126(11):1199–1219, 2018. 3

[32] 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. 1, 3, 7

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 3

[34] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, volume 1, page 4, 2016. 1

[35] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6924–6932, 2017. 1

[36] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 3

[37] Wenjing Wang, Jizheng Xu, Li Zhang, Yue Wang, and Jiaying Liu. Consistent video style transfer via compound regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12233–12240, 2020. 3

[38] Wenjing Wang, Shuai Yang, Jizheng Xu, and Jiaying Liu. Consistent video style transfer via relaxation and regularization. *IEEE Trans. Image Process.*, 2020. 3

[39] Xin Wang, Geoffrey Oxholm, Da Zhang, and Yuan-Fang Wang. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5239–5247, 2017. 1

[40] Hao Wu, Zhengxing Sun, and Weihang Yuan. Direction-aware neural style transfer. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1163–1171, 2018. 1

[41] Xinxiao Wu and Jialu Chen. Preserving global and local temporal consistency for arbitrary video style transfer. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1791–1799, 2020. 3

[42] Zhijie Wu, Chunjin Song, Yang Zhou, Minglun Gong, and Hui Huang. Efanet: Exchangeable feature alignment network for arbitrary style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12305–12312, 2020. 1

[43] Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. Attention-aware multi-stroke style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1467–1475, 2019. 1, 2, 3

[44] Hang Zhang and Kristin Dana. Multi-style generative network for real-time transfer. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1