Consistent Video Style Transfer via Relaxation and Regularization

Wenjing Wang, *Graduate Student Member, IEEE*, Shuai Yang, *Member, IEEE*, Jizheng Xu, *Senior Member, IEEE*, and Jiaying Liu, *Senior Member, IEEE*

***Abstract*— In recent years, neural style transfer has attracted more and more attention, especially for image style transfer. However, temporally consistent style transfer for videos is still a challenging problem. Existing methods, either relying on a significant amount of video data with optical flows or using single-frame regularizers, fail to handle strong motions or complex variations, therefore have limited performance on real videos. In this article, we address the problem by jointly considering the intrinsic properties of stylization and temporal consistency. We first identify the cause of the conflict between style transfer and temporal consistency, and propose to reconcile this contradiction by relaxing the objective function, so as to make the stylization loss term more robust to motions. Through relaxation, style transfer is more robust to inter-frame variation without degrading the subjective effect. Then, we provide a novel formulation and understanding of temporal consistency. Based on the formulation, we analyze the drawbacks of existing training strategies and derive a new regularization. We show by experiments that the proposed regularization can better balance the spatial and temporal performance. Based on relaxation and regularization, we design a zero-shot video style transfer framework. Moreover, for better feature migration, we introduce a new module to dynamically adjust inter-channel distributions. Quantitative and qualitative results demonstrate the superiority of our method over state-of-the-art style transfer methods. Our project is publicly available at: https://daooshee.github.io/ReReVST/.**

***Index Terms*— Style transfer, video processing, image synthesis, temporal consistency, temporal regularization.**

I. Introduction

ARTISTIC images are visually attractive and impressive. In the past, creating artistic imagery used to take experts hours of effort. Now with the technique of style transfer, real scene images can be automatically converted into artistic works, benefiting users without professional capacities. Gatys *et al.* [1] first proposed to use Convolutional Neural Networks (CNNs) for rendering artistic effects, which is referred as Neural Style Transfer (NST). Since [1] rendered images in an iterative optimization way, which is of limited time efficiency, a variety of approaches fastened the stylization process for single-style-per-model [2], multi-style-per-model [3], and zero-shot style transfer [4]. Some researches focused on extending NST for photorealistic rendering [5], doodle style transfer [6], and stereoscopy [7].

With the fast development of the internet and mobile devices, an increasing number of videos are captured and shared. Applying style transfer to video is interesting yet challenging. One difficulty is to maintain temporal consistency for video style transfer. Ruder *et al.* [8] first proposed an online image optimization-based method. However, it takes several minutes to process a single frame even with pre-computed optical flows. To speed up the process, feed-forward models were latter proposed [9]–[11], where picture pairs with optical flows are used to train the network with a temporal consistency loss. Although these models are much faster, their performance on temporal consistency is not comparable with [8].

In addition to speed and performance, existing video style transfer models also have the following problems: 1) Current researches still follow the general solution of training on videos or referring to optical flows, without considering the properties of stylization. 2) Researches [9] have pointed out that there is a trade-off between stylization and temporal consistency. However, due to the inaccuracy of optical flows or the defective design of regularizations, existing methods have an unsatisfactory trade-off rate. 3) One bottleneck of style transfer lies in feature migration. Existing end-to-end trainable feature migration modules are either computationally costly or unable to fully reconstruct style patterns.

In this article, we address the above problems and propose a consistent video style transfer framework, as shown in Fig. 1. First, we point out that the difficulty of video style transfer lies in the contradiction between stylization and temporal consistency, and further put forward to ease this conflict by adjusting the style loss. Through relaxing the shape-sensitivity of the style loss, the temporal consistency can be directly improved. Second, through theoretical analysis, we find that both the traditional way of training with videos and some recently proposed single-frame regularizations have contradictions with the essence of temporal consistency, which can lead to under-fitting and thus degrades network performance. Based on mathematical modeling, we derive a new compound

Manuscript received March 11, 2020; revised July 28, 2020 and August 18, 2020; accepted August 30, 2020. Date of current version September 25, 2020. This work was supported in part by the National Natural Science Foundation of China under Contract 61772043 and in part by the Beijing Natural Science Foundation under Contract 4192025 and Contract L182002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yi Yang. *(Corresponding author: Jiaying Liu.)*

Wenjing Wang, Shuai Yang, and Jiaying Liu are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100080, China (e-mail: daooshee@pku.edu.cn; williamyang@pku.edu.cn; liujiaying@pku.edu.cn).

Jizheng Xu is with ByteDance Inc., San Diego, CA 92122 USA (e-mail: xujizheng@bytedance.com).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TIP.2020.3024018

Fig. 1. On the left are input style images, and on the right are three frames from corresponding videos. The first row is the input video, the second and third rows show our video style transfer results, and the fourth row is our result of making the style dynamically change from one style image to another.

regularization to better fit the nature of temporal variation. Extensive experiments demonstrate the effectiveness of the proposed regularization in balancing temporal stability and stylization effect. Another aspect to improve video style transfer is to introduce the inter-frame relationship, which benefits long-term temporal consistency. However, many models implement this by estimating optical flows, which is of limited robustness and low efficiency. We instead propose to share global features. With feature distributions consistent among the whole sequence, networks become more robust to motions and illumination changes without hurting the stylization effect. Finally, for feature migration, we design a dynamic filter to adjust inner- and inter-channel feature distributions, which can better reconstruct style patterns.

Combing the above improvements, we present a video style framework. Our contributions summarize as follows:

- We propose a novel video style transfer framework with both superior temporally consistency and visually pleasing stylization effect. A powerful end-to-end trainable filter is developed for dynamically adjusting inter-channel feature distributions, which improves color and texture reconstruction. Experimental results demonstrate the superiority of the proposed framework.
- We analyze the conflict between stylization and temporal consistency. To alleviate the conflict, we attempt to relax the style loss through well-designed motions, which can effectively improve temporal consistency without degrading style transfer effects.
- We provide a theoretical analysis and universal modeling of temporal consistency, from which we derive a novel regularization which has superior effectiveness in balancing spatial and temporal performance, and can help with other computer vision video tasks.

Note that, this article is an extension of our earlier publication [12]. Our changes can be summarized into three aspects: method improvement, experiment enrichment, and prospect exploration. First, this article proposes a new technique for video style transfer. In addition to [12], we analyze the temporal stability of stylization in Sec. III, and resolve the conflict between stylization and temporal consistency through relaxing the objective function in Sec. IV. With the new technique, we achieve a performance higher than our original submission. Second, we conduct more experiments and related analysis. We provide additional comparison results on special cases of changing focus, changing illumination, and large moves. Ablation studies explore the separate effects of our proposed techniques. Also, we show the application of style interpolation and cases of degraded videos. Third, we explore to combine the proposed techniques with Generative Adversarial Networks (GAN), which might inspire new directions for future research.

The rest of the paper is organized as follows. In Sec. II, we review existing researches on style transfer and temporal consistency. After that, in Sec. III, we provide modeling of temporal consistency, and analyze the temporal stability of style transfer. Then, in Sec. IV, we propose solutions for maintaining temporal consistency, which can solve the problems we find in Sec. III. After that, in Sec. V, we introduce our full framework for consistent video style transfer. Later, in Sec. VI, we conduct experiments to demonstrate the superiority of the proposed method. We provide extensive comparison results, ablation studies, and applications. Finally, in Sec. VII, we draw conclusions and discuss potential future research directions.

II. Related Works

A. Image Style Transfer

Style transfer is the task of migrating styles from an artistic image to a target image. Neural Style Transfer (NST) [1] first formulated style as the feature-level correlation of pre-trained image classification convolutional neural networks. Since then, stylization has received more and more attention, even gives birth to industrial products such as Prisma[1] and DeepArt.[2]

A lot of research has been done to accelerate NST. Johnson *et al.* [2] turned online optimization into feed-forward processing by training a network with content loss and style loss. Johnson *et al.*'s model can only remember one style at a time. Chen *et al.* [3] proposed to store styles in residual blocks. Recent researches mainly focus on rendering images to any style in one feed-forward pass. Following the idea that the essence of style transfer is to migrate feature distributions [13], most zero-shot methods designed feature adaptation modules. AdaIN [4] adjusted features through mean and variance. WCT [14] proposed multi-scale whitening and coloring transformation. Chen and Schmidt [15] migrated features by patch swapping. Avatar-Net [16] adopted a style-swap based style decorator. Recently, Li *et al.* [17] designed a linear transformation matrix. SANet [18] proposed to replace the patch-based mechanism with a linear module. Yao *et al.* [19] introduced a self-attention mechanism.

Existing image style transfer methods have no consideration of temporal consistency, therefore result in severe flickering artifacts on videos. In this article, a novel style transfer framework is proposed, which performs better for both temporal

[1] https://prisma-ai.com/
[2] https://deepart.io/

consistency and style transfer effects. Moreover, existing modules for feature migration are either computationally costly such as WCT [14], unable to fully render the artistic style such as AdaIN [4], or not end-to-end trainable such as Avatar-Net [16]. To resolve these problems, we design a dynamic filter for inter-channel feature adjustment.

B. Video Style Transfer

Video stylization approaches can be divided into two categories: multiple-frame and single-frame.

Multiple-frame-based methods consider inter-frame correlation in the inference phase. Based on NST, Ruder *et al.* [8] warped previous frames to the current time, which forms a temporal loss to guide the optimization. Based on Split and Match [20], Frigo *et al.* [21] also used optical flows. For further acceleration, some feed-forward networks are proposed [10], [11], [22]–[24]. Gupta *et al.* [11] changed the input of [2] into a concatenation of the current content and the warped previous stylization result. Chen *et al.* [10] proposed to do the warping in the feature space.

The effect of multiple-frame methods highly depends on the correctness of estimated inter-frame correlation, such as optical flows or RNNs. Therefore, ghosting artifacts may occur when the estimation is inaccurate. Moreover, either estimating optical flows or using RNNs is computationally expensive, making it impractical to process high-resolution or long videos. Besides, the spatial distribution of style patterns is often neglected, which may lead to weird results.

Single-frame-based models instead process each frame independently. The ability to maintain temporal consistency is usually obtained through training loss functions [9] or stable modules [17]. However, Huang *et al.* [9] required a separate network for each style, while the stylization effects of Linear [17] is not satisfactory.

For multi-frame methods, introducing inter-frame information helps with maintaining temporal consistency. However, as mentioned before existing forms of inter-frame information have many drawbacks. Single-frame methods are robust to motions and computationally effective. However, existing techniques are not powerful enough. In this article, we propose a framework that combines the advantages and avoids the shortcomings of these two kinds of methods.

C. Temporal Consistency

Temporal consistency is an important factor for video quality. Optical flow is the most common tool for maintaining temporal consistency [8], [10], [11], [22]. However, estimating optical flow is computationally expensive. Temporal filtering has been used in some researches [25]–[27], however, the designed filter formulation is specific to the target task and cannot be easily generalized to other applications.

Some techniques target universal tasks. Bonneel *et al.* [28] proposed a gradient-domain technique that is blind to the particular image processing algorithm. Yao *et al.* [29] targeted at the cases of occlusion and proposed an online keyframe strategy and a local color affine model. However, these approaches cannot handle more complicated tasks such as style transfer. Lai *et al.* [30] designed a blind post-processing neural network that supports many kinds of vision tasks. However, for style transfer, it may blur the strokes and bring out a color cast. Eilertsen *et al.* [31] proposed to use single-frame regularization to increase the temporal stability of CNNs. However, there is no theoretical basis and no experimental comparison against existing video models.

Combining style transfer, in this article we further study the theoretical principle of temporal consistency, propose an effective solution, and conduct extensive experiments to demonstrate our standpoints.

III. Temporal Properties of Style Transfer

In this section, we provide modeling for temporal consistency, based on which we analyze the conflict between stylization performance and temporal consistency.

A. Modeling of Temporal Consistency

Different from separate images, video frames are highly correlated and temporally smooth. To analyze the flicker in video style transfer, we first need to give a mathematical formulation of this inter-frame correlation.

Without loss of generality, we ignore the appearance and disappearance of objects and assume that the color is constant. In this way, we only need to consider motions. Denote X_n as the n-th frame of the video, temporal consistency can be defined as: there exists a small number $\delta > 0$, such that for all n, m with $|n - m| < K$, the value of X_n satisfies

$$||X_n - W_{X_m \to X_n}(X_m)|| < \delta, \tag{1}$$

where K denotes the length of long-term temporal consistency, and $W_{X_m \to X_n}(X_m)$ denotes the result of warping X_m to X_n with the corresponding optical flow. δ defines the degree of consistency. For a stable video, δ should be small enough so that human eyes are not sensitive to the flickering artifacts.

Our target of maintaining the temporal consistency of an image processing model $\mathcal{F}$ can be expressed as: when the input video is consistent, we hope that the output video is also consistent. Evidently, X_n and X_m are not limited to adjacent video frames. Therefore, we can write our target as

Target 1: There exists small numbers ϵ, $\delta > 0$, such that for any X, X', and an operation of warping W with $||X' - W(X)|| < \delta$, $Y = \mathcal{F}(X)$ and $Y' = \mathcal{F}(X')$ satisfy

$$||Y' - W(Y)|| < \epsilon. \tag{2}$$

B. Temporal Consistency of Style Loss

Many image processing models [32]–[35] have little flickering artifacts on videos. But for image style transfer, the contrary is true. Existing researches [10], [11], [22]–[24] have found that image style transfer models are not temporally consistent. Gupta *et al.* [11] explained that this is due to the non-convexity of the Gram-Matrix-based minimization objective. However, many variants of the style loss do not use the Gram Matrix but still cannot get rid of the problem. In this article, we offer a more general explanation.

TABLE I

STYLE LOSS WITH AND WITHOUT THE OPTIMIZED MOTION W. WE ALSO SHOW THE MEAN RATIO OF $\mathcal{L}_{style}(W(Y))/\mathcal{L}_{style}(Y)$

Network Φ	Distance $D(\cdot)$	$\overline{\mathcal{L}}_{style}(Y)$	$\overline{\mathcal{L}}_{style}(W(Y))$	Ratio
VGG16	Gram	6.451	7.928	1.508
VGG16	Mean & Var	3.008	4.127	1.518
VGG19	Gram	22.886	25.094	1.263
VGG19	Mean & Var	4.294	5.742	1.497
Average				1.446

The task is to render the style of an artistic image S to a target image I. Without loss of generality, we assume that the style image is fixed during the training. Although there are many variants, existing style losses have not gone beyond characterizing the difference between the distributions of two image features as follows,

$$\mathcal{L}_{style}(I) = \sum_l (D(\Phi_l(I) - \Phi_l(S))), \tag{3}$$

where Φ_l is the l-th layer of a pre-trained model Φ. $D(\cdot)$ is the distribution distance, such as the Gram Matrix [1], mean and variance [13], and the Earth Movers Distance [36], *etc*. In this article, we choose the mean and variance version for its low computation cost. The loss formulation is

$$\mathcal{L}_{style}(I) = \sum_l (||\text{Mean}(\Phi_l(I)) - \text{Mean}(\Phi_l(S))||^2 + ||\text{Var}(\Phi_l(I)) - \text{Var}(\Phi_l(S))||^2), \tag{4}$$

where Mean$(\cdot)$ denotes calculating the mean, and Var$(\cdot)$ denotes calculating the variance.

Different from [11], we hold that the temporal instability of the style loss comes from the features Φ_l. To perceive the lines, strokes, and colors of the artistic image, Φ has to be sensitive to the variation of shapes and edges, such as VGG [37]. However, temporal consistency instead requires the model to be robust to the variation of shapes and edges, indicating that there is a conflict between style transfer and temporal consistency.

To demonstrate this conflict, we show that particular motions can greatly increase the style loss. We first selected 6 content images and 20 artistic images, then applied 7 style transfer methods [1], [4], [14], [16]–[19] and obtained 840 style transfer results Y. Next, we used Stochastic Gradient Descent (SGD) to find an optical flow W to maximize $\mathcal{L}_{style}(W(Y))$. An example is shown in Fig. 2. We limited W to be small so that Y and $W(Y)$ are very similar in human eyes. To cover all commonly-used types of style losses, we tested on both VGG-16 and VGG-19. For distance metrics, we used not only the original version, *i.e*. the Gram Matrix, but also a widely-used variant of combining mean and variance [13].

In Fig. 2, although Y and $W(Y)$ absolutely share the same artistic style in human eyes, the style loss of $W(Y)$ is almost twice as much as that of Y. As shown in Table I, for all different kinds of Φ and distance metrics, we can easily find a W to make the style loss rise by 44.6%.

Assume that $Y = \mathcal{F}(X)$ is a perfect style transfer output, we can always find some W with which the style loss of $Y' = \mathcal{F}(W(X))$ is pretty high. Combining Eq. (2), this means that the style loss and temporal loss might not be well minimized at the same time. Therefore, image stylization models that target at minimizing style loss cannot avoid flickering artifacts.

(a) Inputs (b) Y (c) $W(Y)$ (d) W

Fig. 2. Style loss can vary a lot with a small motion. Stylization result Y is obtained by NST from the inputs. We use the optical flow W to warp the image Y, and obtain $W(Y)$. The style loss of Y is 0.77, while the style loss of $W(Y)$ is 1.32.

The above analysis also indicates that if the stylization model is forced to maintain temporal consistency, the stylization performance will degrade. Corresponding experimental results can be found in [9]. Huang *et al.* trained the network of [2] with both stylization loss and temporal loss for video style transfer. Although the model is temporally smoother, the distribution of color is simpler, and the details of strokes and textures are lost.

IV. MAINTAINING TEMPORAL CONSISTENCY VIA TRAINING ON SINGLE-FRAME

In this section, we introduce a comprehensive solution for video style transfer. First, to reduce the conflict between stylization and temporal consistency, we relax the style loss. Then, a new temporal regularization is derived from the temporal consistency modeling. For the long-term consistency, we propose a strategy of sharing global features.

A. Relaxed Style Loss

Style loss is sensitive to shape variation, but in human eyes, shape may not play an important role in defining artistic style. As previously mentioned, slightly warping the stylization result Y has an effect on style loss. We further find that the same is true for the style image S. In Fig. 3, although the two style images are of the same style in human eyes, using Fig. 3(b) as S, the style loss is 0.7479, while using Fig. 3(c) as S, the style loss is only 0.5711.

This inspires us to remove the shape-related parts in style loss, which means to relax the style loss so that it will be less sensitive to motions. Toward this end, we find a motion W to decrease the style loss

$$\mathcal{L}^*_{style}(Y, S) = \min_W \mathcal{L}_{style}(Y, W(S)). \tag{5}$$

If we instead adjust the stylization result Y

$$\mathcal{L}'^*_{style}(Y, S) = \min_W \mathcal{L}_{style}(W(Y), S), \tag{6}$$

the style loss will also be decreased. However, in experiments, we found that adjusting Y can even increase the temporal loss.

(a) Image (b) Style A (c) Style B (d) Residual

Fig. 3. Computing style loss using different style images. (a) a content image and its artistic version, (b)-(c) style images, (d) the residual of (b) and (c). Using style image (b), the style loss of (a) is 0.7479, while using style image (c), the style loss of (a) is 0.5711.

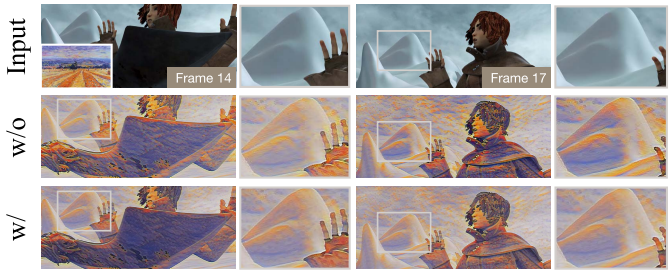


Fig. 4. Ablation study of global feature sharing. With the proposed strategy, although a black broadsword enters the picture, stylized patterns of the snow mountain maintain the same appearance on different frames.

This may be because $W(\cdot)$ distorts the output and introduces random factors, making the problem more difficult.

Since there is no analytical solution, we solve the minimization problem through optimization. To restrict solution space and smooth the optical flow, we transform W into

$$W = \text{Blur}(\lambda_{wm} \tan(\text{Upsample}(W_s))), \tag{7}$$

where W_s is 1/8 smaller than W, $\tan(\cdot)$ denotes the tangent function, and $\text{Blur}(\cdot)$ is a Gaussian blur operation of standard deviation value $\sigma = 50.5$. The kernel size is set to a very high value 101 so that W can be smooth enough. We use tan and λ_{wm} to limit the range of the motion.

To fasten the training process, we use a very large learning rate of $lr = 16$ and limit the iteration to 16 steps only. Finally, the proposed relaxed style loss is

$$\mathcal{L}^*_{style}(Y, S) = \min_{W_s} \mathcal{L}_{style}(Y, W(S)). \tag{8}$$

B. Compound Temporal Regularization

Existing video-related models [9] are trained on a large amount of videos with corresponding optical flows. However, high-quality videos can be difficult to collect. Moreover, state-of-the-art optical flow estimation models are still not good enough, which may lead to inaccurate labels and mislead the style transfer model. In this article, we design a new temporal regularization based on our temporal consistency modeling, with which no video or estimated optical flow is required in the training process.

Denote $\Delta = X' - W(X)$, then Eq. (2) is

$$\begin{aligned} ||Y' - W(Y)|| &= ||\mathcal{F}(X') - W(\mathcal{F}(X))|| \\ &= ||\mathcal{F}(W(X) + \Delta) - W(\mathcal{F}(X))||, \end{aligned} \tag{9}$$

which means that maintaining temporal consistency is equivalent to minimizing

$$\mathcal{L}_{comp} = ||\mathcal{F}(W(X) + \Delta) - W(\mathcal{F}(X))||. \tag{10}$$

Intuitively, $\mathcal{L}_{comp}$ is a compound of two transformations: Δ represents local jitter or noise, while $W(\cdot)$ represents motions. The explicit meaning of Eq. (10) is to force the model to generate consistent results under the compound transformation.

One difference between the classic way of training with video and our regularization is that, for the former, users first obtain adjacent frames and then estimate optical flows, where the estimation may be inaccurate. For our regularization, users first generate optical flows and then synthesize adjacent frames. Therefore the optical flows are absolutely accurate. We will show later that training with $\mathcal{L}_{comp}$ can efficiently improve temporal consistency.

C. Sequence-Level Global Feature Sharing

Single frame information is obviously insufficient for stable video processing. A common way to introduce inter-frame correlation is warping frames with optical flows in the inference phase [8]. However, this highly relies on the accuracy of optical flows and fails to handle long-term consistency.

We notice that many style transfer methods use full image global distributions to characterize styles, such as feature-level mean and variance in AdaIN [4]. However, when there are extreme variations, *e.g.* a new object enters or the illumination changes, global distributions will be changed. This may cause the same object to have different styles in different frames.

Driven by this observation, we propose to share global features across the whole sequence. Specifically, we first extract 1/8 frames, then calculate the sequence-level average of the global features. Finally in the inference phase, only the averaged values are used.

As shown in Fig. 4, without the sequence-level global feature sharing, stylized patterns are of different appearances in different frames, which creates flicker artifacts.

V. Consistent Video Style Transfer

Combining the above techniques for temporal consistency, we propose a novel video style transfer framework. In this section, we first introduce our new stylization module for adjusting feature distribution. Then, we give the architecture of our full model.

A. Dynamic Inter-Channel Filter

According to [13], the essence of style transfer is to migrate feature distributions. Currently, most feed-forward arbitrary style transfer models are based on feature adjusting modules. AdaIN [4] proposed to align the feature-level channel-wise mean and variance and designed the Adaptive Instance Normalization (AdaIN) layer. Avatar-Net [16] proposed the Style Decorator, combining distribution migration and patch matching. Yao *et al.* [19] improved Avatar-Net with self-attention mechanisms. Also based on the self-attention mechanism,

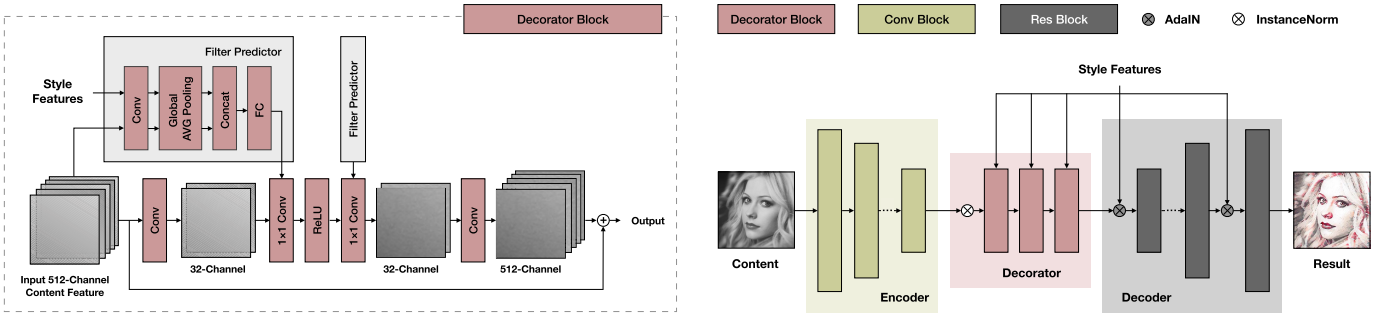


Fig. 5. Left: the proposed decorator block for inter-channel feature adjustment. Both target style features and input content features are fed into a shallow sub-network Filter Predictor to predict filters. Residual learning and dimensionality reduction are used to improve the efficiency. Right: The overall architecture of the proposed encoder-decoder style transfer network.

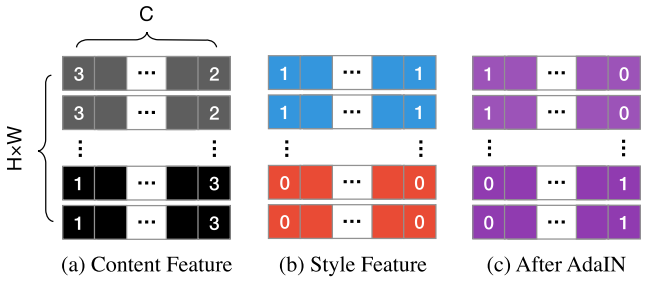


Fig. 6. AdaIN edits the feature of every single channel but has no consideration of the correlation between different channels.

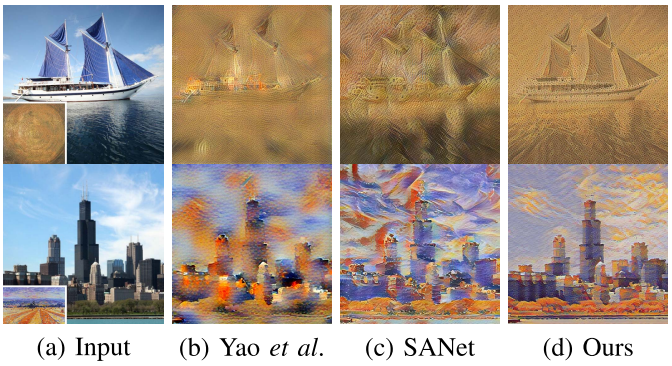


Fig. 7. Comparison with other style transfer modules.

SANet [18] transferred features by modifying the attention. However, as shown in Fig. 7, AvatarNet and [19] cannot well match style patterns and semantic structures, which may lead to messy textures and distorted contours. Moreover, AvatarNet does not support end-to-end training. SANet may distort textures and introduce strange strokes. To solve these problems, we propose a new style transfer module.

Our design is based on AdaIN. In AdaIN, although for every single channel the distributions are well migrated, the correlation across different channels may be still inconsistent with that of the target style, as illustrated in Fig. 6. This can lead to unsatisfactory results, such as fusing colors in Fig. 8(b). Based on this observation, we propose to adjust both inner- and inter-channel features.

As shown in Fig. 5, both input and style features are fed into the Filter Predictor module to dynamically predict a linear combination of different channels, which is later applied to the

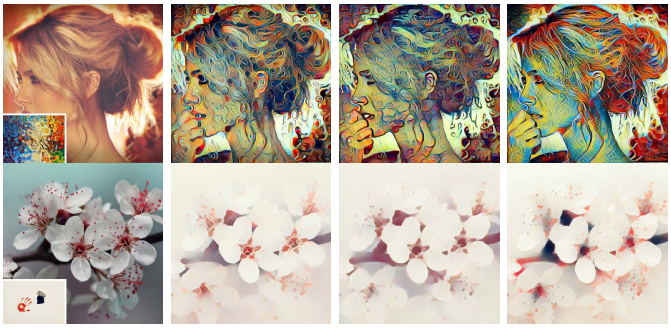


Fig. 8. Ablation study of different decorator modules. S denotes being dynamic to only style features. $C + S$ denotes being dynamic to both content and style features.

input feature by a 1×1 convolution. Global average pooling modules guarantee that the network is robust to any resolution. Notice that if we directly predict a 512-channel filter, which has $512 \times 512 = 2^{18}$ parameters, the computational complexity will be too high. Therefore, we reduce the dimension to 32 and use residual learning to prevent information loss. As shown in Fig. 8, with Filter Predictor, the textures are clearer and the colors match the target style better.

The proposed Filter Predictor is dynamic to both content and style. If it is only dynamic to style, which is similar to Meta Network [38], the stylization result will be unsatisfactory as shown in Fig. 8(c).

B. Network Architecture

Our model is based on the widely-used encoder-decoder architecture. It has shown great performance in many applications, such as segmentation [39], compression [40], unsupervised feature learning [41], and also style transfer [4], [14], [19]. Common encoder-decoder architectures contain one encoder for feature extraction and one decoder for image generation. While in our model, there are two encoders, one for the content and the other for the style. Also, there is a feature migration module between the encoder and the decoder.

The architecture of the proposed framework is shown in Fig. 5. First, VGG-like encoders extract high-level features of the input content and style images, respectively. Then,

Fig. 9. Results with and without content image desaturation. (a) and (c) are the inputs for (b) and (d), respectively. In (b), the model is trained without content image desaturation, and (d) vice versa.

the content features are processed by three consecutive decorator blocks to match the styles. Finally, the adjusted features are mapped back to the image domain through a decoder. Inspired by Avatar-Net [16], we add multi-level AdaIN skip connections between the encoder and the decoder.

In our model, there are two kinds of global features: 1) the feature-level mean and variance in instance normalizations or AdaINs; 2) filters predicted by Filter Predictor. According to the proposed global feature sharing strategy, their values are shared among the whole sequence in the inference phase.

C. Loss Functions

The training loss $\mathcal{L}$ consists of five functions:

$$\mathcal{L} = \lambda_t \mathcal{L}_t + \lambda_s \mathcal{L}_s + \lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r + \lambda_{tv} \mathcal{L}_{tv}, \tag{11}$$

where $\mathcal{L}_t$ denotes the proposed compound temporal loss, and $\mathcal{L}_s$ denotes the relaxed style loss. For style loss $\mathcal{L}_s$ and content loss $\mathcal{L}_c$, we use a pre-trained VGG-19 [37]:

$$\mathcal{L}_s(S, Y) = \min_{W_s} \tilde{\mathcal{L}}_s(W(S), Y), \tag{12}$$

$$\tilde{\mathcal{L}}_s(S, Y) = \sum (||\text{Mean}(\Phi_l(S)) - \text{Mean}(\Phi_l(Y))||^2 + ||\text{Var}(\Phi_l(S)) - \text{Var}(\Phi_l(Y))||^2), \tag{13}$$

$$\mathcal{L}_c(C, Y) = \sum ||\Phi_l(C) - \Phi_l(Y)||^2, \tag{14}$$

where Φ_l denotes the feature map of VGG-19 at layer l, S denotes style images and C denotes content images. For $\mathcal{L}_c$, we use *ReLU4_1*. For $\mathcal{L}_s$, we use *ReLU1_1*, *ReLU2_1*, *ReLU3_1*, and *ReLU4_1*. $\mathcal{L}_{tv}$ denotes total variation loss [42].

Stylization can be easily affected by the color of content images, especially when the style image is blue and the content image has red or yellow objects. As shown in Fig. 9(b), the walls are red and the lighthouse is orange, which are inconsistent with the blue color of the style images. This may be due to that content loss use image classification models to extract features and these models are sensitive to warm colors. To solve this problem, we desaturate content images in both training and inference phases. The final content loss for training is:

$$\mathcal{L}_c(C, Y) = \sum ||\Phi_l(C_{gray}) - \Phi_l(Y)||^2, \tag{15}$$

where C_{gray} denotes gray content images. As shown in Fig. 9, desaturation helps with preserving the color of the style.

Moreover, we introduce a color reconstruction loss:

$$\mathcal{L}_r = ||\mathcal{F}(C_{gray}, C_{color}) - C_{color}||, \tag{16}$$

where $\mathcal{F}$ represents our model, and $\mathcal{F}(C_{gray}, C_{color})$ denotes colorizing gray images using the style of the corresponding colorful images. The color reconstruction loss helps with improving the temporal consistency and avoiding color bias.

VI. Experimental Results

A. Implementation Details

In compound regularization, $W(\cdot)$ is implemented by warping with a random optical flow. For a frame of size $H \times W$, first a Gaussian map W_{tw} of shape $H/100 \times W/100 \times 2$, mean 0, and standard deviation σ_w is generated. Then, W_{tw} is upsampled to the size of $H \times W$ and blurred by a Gaussian filter of kernel size 100. Through upscaling and blurring, W_{tw} would be very smooth. Finally, we add two random values W_{tl} of range $[-10, 10]$ to the Gaussian map, and obtain the final motion $W = W_{tw} + W_{tl}$. Intuitively, W_{tw} represents wavy twists and W_{tl} represents translational motion.

The other transformation Δ is a random noise with $\Delta \sim \mathcal{N}(0, \sigma^2 I)$, $\sigma \sim \mathcal{U}(\sigma_s, 2\sigma_s)$. We use σ_w and σ_s to control the transformation magnitude.

The final network is first pre-trained with the original style loss $\tilde{\mathcal{L}}_s$ and without temporal loss $\mathcal{L}_t$ for two epochs, then fine-tuned with the relaxed style loss $\mathcal{L}_s$ for one epoch, finally fine-tuned with $\mathcal{L}_s$ and the temporal loss $\mathcal{L}_t$ for one epoch. More experimental settings can be found in the supplementary material. The whole training process takes about 8 hours on GeForce RTX 2080Ti. Using relaxed style loss can triple the time of each iteration, but does not affect the testing time.

B. Quantitative and Qualitative Comparisons

The proposed method is compared with five state-of-the-art style transfer frameworks [1], [4], [8], [14], [16], [17]. We also compared with using the post-processing model [30] with [14] and our baseline model. Our baseline denotes the style transfer network with the original style loss, and without temporal loss and global feature sharing.

1) Temporal Consistency: For quantitative comparison, we employ the widely used temporal loss [8]. We evaluated both short and long-term temporal consistency:

$$\mathcal{L}_{lt} = ||O \circ (W_{X_n \to X_{n-i}}(Y_n) - Y_{n-i}))||, \tag{17}$$

where $i \in \{1, 2, 4, 8, 16\}$ denotes frame interval, O denotes occlusion mask, and $\circ$ denotes element-wise multiplication. We used all the sequences of MPI Sintel dataset [43]. For $i = 1$, MPI Sintel provides ground truth optical flows. For $i > 1$, we used PWC-Net [44] to estimate optical flows. Since optical flows might be inaccurate, we modified occlusion mask as

$$O' = O \cup \{||W_{X_n \to X_{n-i}}(X_n) - X_{n-i})|| > 10\}. \tag{18}$$

TABLE II

QUANTITATIVE EVALUATION OF TEMPORAL CONSISTENCY. OUR MODEL YIELDS THE LOWEST TEMPORAL LOSS FOR ALL TEMPORAL LENGTH

Method	Temporal Loss / Interval i				
	$i = 1$	$i = 2$	$i = 4$	$i = 8$	$i = 16$
WCT	0.116	0.119	0.120	0.116	0.112
AdaIN	0.082	0.085	0.087	0.086	0.085
WCT + Blind	0.070	0.073	0.077	0.080	0.083
Avatar-Net	0.056	0.063	0.067	0.070	0.073
Linear	0.040	0.046	0.049	0.051	0.053
Ruder *et al.*	0.038	0.047	0.059	0.073	0.086
Baseline	0.059	0.062	0.063	0.063	0.063
Baseline + Blind	0.048	0.050	0.052	0.056	0.063
Ours	**0.036**	**0.040**	**0.042**	**0.043**	**0.044**

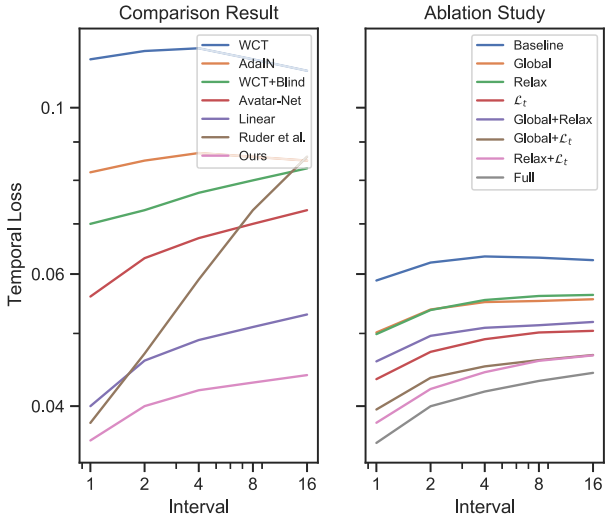


Fig. 10. Visualization of temporal consistency. Left is the comparison results against other methods. Right is the ablation study of our model.

This also helps us exclude areas where the illumination changes. For styles, we collected 20 artworks of various types.

Quantitative results are shown in Table II and Fig. 10. The model proposed by Ruder *et al.* [8] maintains good consistency for the short temporal length, however, with the increase of i, the performance degrades heavily. This is because it relies too much on the inter-frame relationship and can be easily affected by inaccurate optical flows. Lai *et al.* [30] designed a blind post-processing model Blind. We show its result with WCT [14] (which the authors used in the original paper) and our baseline. Blind is also not robust to the increase of temporal length. Compared with other methods, our model yields the best result for all temporal length.

2) Stylization Effects: Image style transfer results are shown in Fig. 11. NST [1] fails to balance colors and leaves many parts not stylized. AdaIN [4] and WCT [14] distort content structures heavily. Avatar-Net [16] well generates the style patterns, however, the patterns have less correlation with the semantic structure. Linear [17] introduces weird colors such as pink in the first row. Our models better balance style migration and semantic reconstruction.

Video style transfer results are shown in Fig. 12. Both AdaIN, Linear, and [8] fail to reconstruct the pure blue color of the target style. WCT and Avatar-Net well synthesize the water painting stroke. However, they large high temporal errors. Compared with other methods, the proposed model better migrates colors and preserves semantic details.

TABLE III

MEAN USER PREFERENCE OF VIDEO STYLE TRANSFER METHODS

Method	Rater Preference		
	Stylization	Temporal	Comprehensive
Ours / WCT	**0.53** / 0.47	**0.95** / 0.05	**0.83** / 0.17
Ours / WCT + Blind	**0.80** / 0.20	**0.98** / 0.02	**0.90** / 0.10
Ours / AdaIN	**0.92** / 0.08	**0.98** / 0.02	**0.97** / 0.33
Ours / Avatar-Net	**0.60** / 0.40	**0.88** / 0.12	**0.77** / 0.23
Ours / Ruder *et al.*	**0.85** / 0.15	**0.92** / 0.08	**0.97** / 0.03
Ours / Linear	**0.73** / 0.27	**0.85** / 0.15	**0.78** / 0.22

Fig. 14 shows the qualitative results of long temporal consistency. For [8] and Blind, the bamboo leaves are stylized differently in frame 1 and 50. Moreover, [8] has a red ghosting artifact of the human's walking track, while Blind causes severe color bias. Our model instead well maintains temporal consistency even for an interval of 49 frames.

In Fig. 13, we show more comparison results against [8]. In Fig. 13(a), the tree becomes more and more blurry as the focus changes. However, in the result of [8], the appearance of the tree is static. Since [8] relies on optical flows and optical flows cannot handle focus, [8] fails to characterize the focus variation. Our model instead uses textures to express focus variation. Similarly, in Fig. 13(b), [8] fails to handle the illumination changes, while our result can reflect the change of shadows on the background. In Fig. 13(c), the result of [8] is messy, making it difficult for the users to recognize the content. Moreover, [8] fails to well transfer the style of the artistic images. In Fig. 13(a), the red color of the apple and some green color of the tree still remain on the stylization result. In Fig. 13(b), [8] fails to generate the brown pencil drawing texture. In Fig. 13(c), [8] has weird white color. In addition, [8] is based on online optimization, taking several minutes to process a single frame even with GPU. Our model can well represent the content variation of the input video while preserving the style of the artistic image, and takes only 0.07 seconds to process a 1024×436 frame.

3) Human Preference: We conduct a user study where 20 people are asked to compare the video style transfer result of the proposed method against other methods. We ask the users to consider visual stylization effect, temporal consistency, and comprehensive effect, respectively. As shown in Table III, our model has the best human visual effect.

4) Exclusion Time: We fix the resolution of the content frames and style images to 512×512, and fix the video sequence length to 50 frames. The hardware devices are GeForce RTX 2080Ti and Intel Xeon CPU E5-2650. We show the average time of generating one frame in Table IV. Due to the slow process of optical flow estimation and optimization-based stylization, [8] takes minutes to generate a single frame. Image style transfer models need to re-encode the style image each time processing a single frame, therefore, the running times of Avatar-Net [16], WCT [14], and AdaIN [4] are all higher than that of our method. Linear [17] is slightly faster

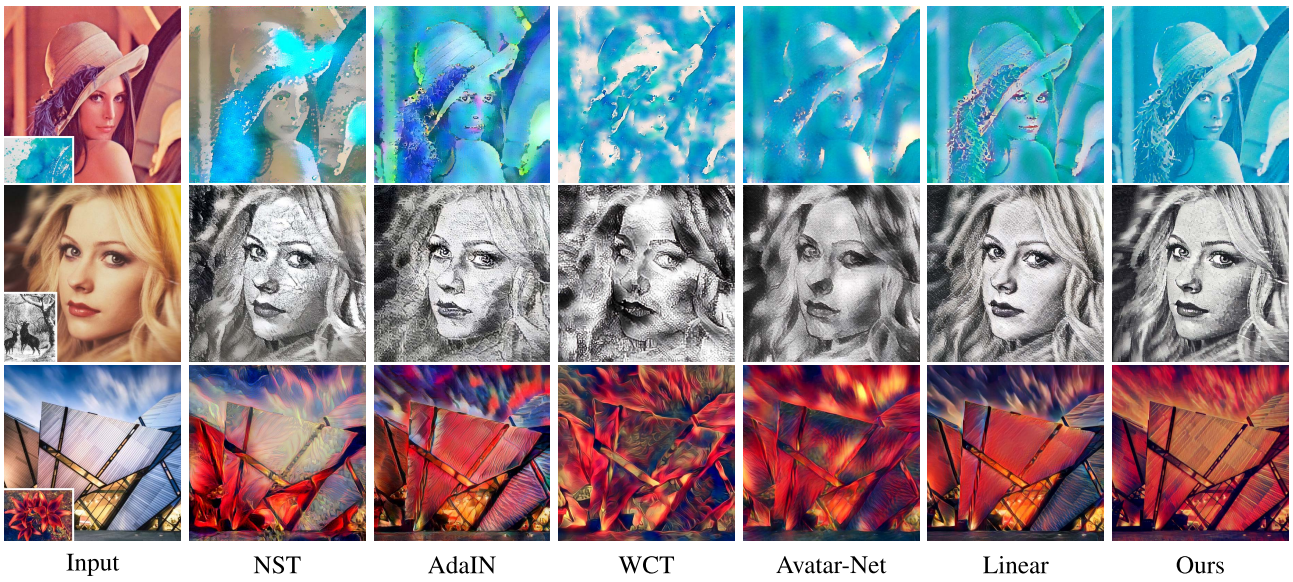


Fig. 11. Comparison with state-of-the-art methods on image style transfer.

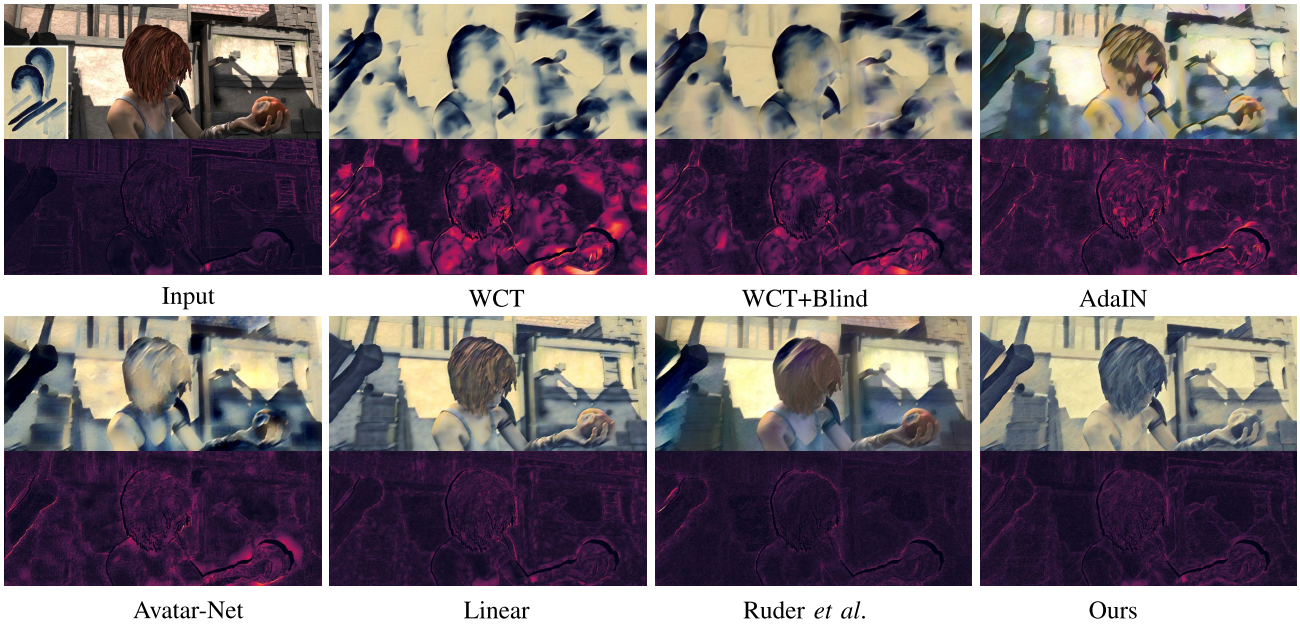


Fig. 12. Comparisons on video style transfer. The bottom of each row shows the temporal error heat map. Please refer to the supplementary materials for a video demonstration.

TABLE IV
EXECUTION TIME COMPARISON (SECONDS)

Method	Ruder *et al.*	WCT	Avatar-Net
Time (sec)	75.320	3.499	2.786
Method	AdaIN	Linear	Ours
Time (sec)	0.050	0.023	0.043

than our model, but as shown in Fig. 10 and Fig. 12, has worse temporal consistency and stylization performance.

C. Ablation Studies

In this section, we evaluate the proposed techniques for temporal consistency: relaxed style loss (Relax), compound regularization ($\mathcal{L}_t$), and global feature sharing (Global).

1) Temporal Consistency: The quantitative results are shown in Table V and Fig. 10. Relaxing style loss and sharing global features have a similar degree of improvement. Their combination can further promote temporal consistency. Using temporal regularization has a significant effect. Finally, using all the techniques yields the best result. As shown in Table V, adding a proposed technique always improves performance, indicating that our three techniques do not interfere with each other and their effects do not overlap. This supports our methodology of comprehensively solving the problem from different perspectives and demonstrates the effectiveness of our modeling.

2) Stylization Effect: Qualitative results are shown in Fig. 17. Sharing global features may slightly change the color distribution, however, the overall artistic effects remain the same. Relaxing style loss even enhances the strokes. Both

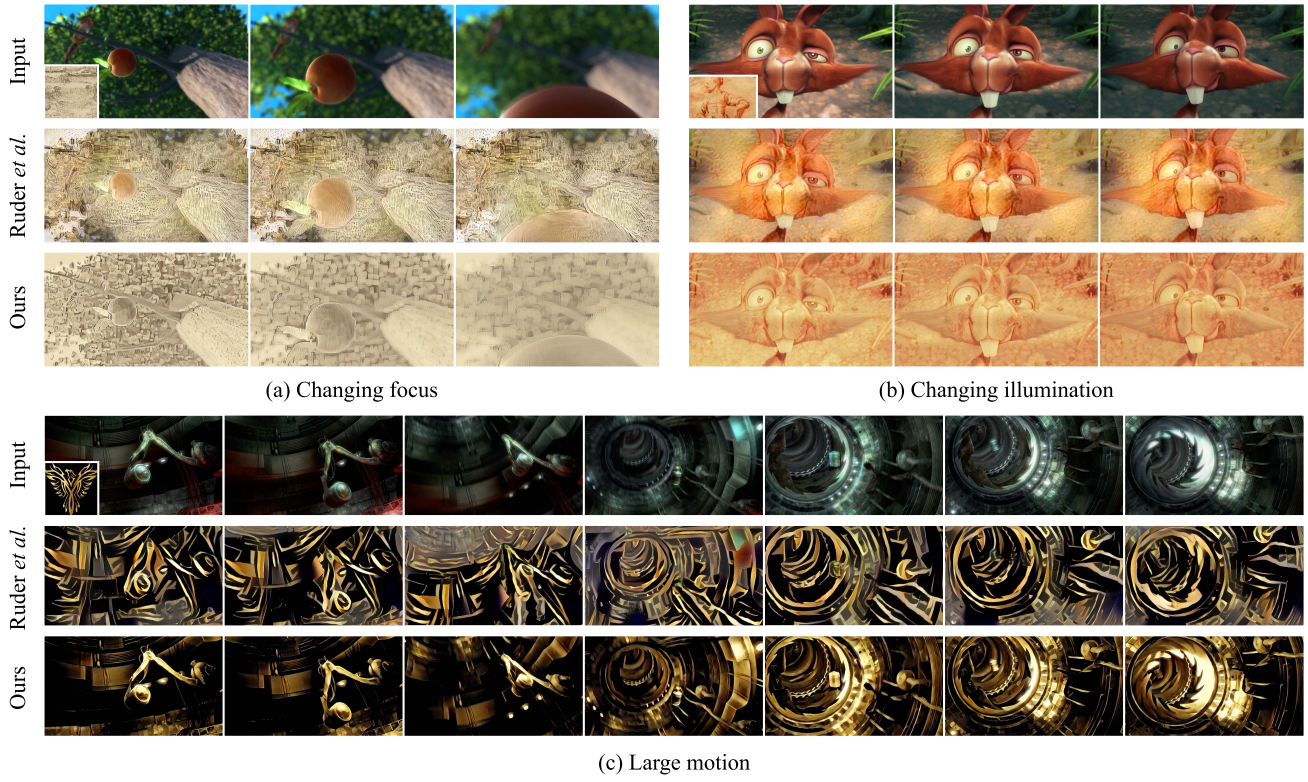


Fig. 13. More comparison results on special cases of (a) changing focus, (b) changing illumination, (c) large motion. Compared with Ruder *et al.*, the result of our model fully reflects the content of the input video. Besides, the texture and color of our result are more consistent with the input style image.

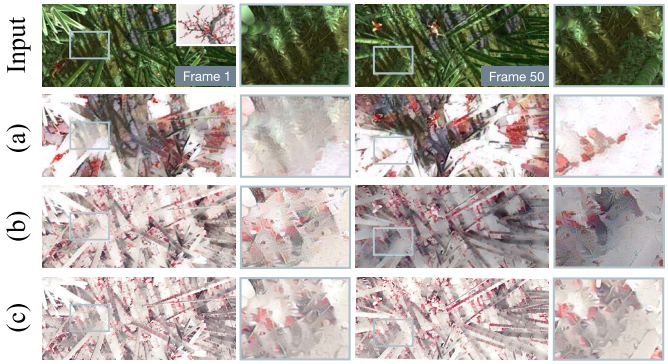


Fig. 14. Comparison results on long-term temporal consistency: (a) Ruder *et al.*, (b) Baseline + Blind, (c) Ours.

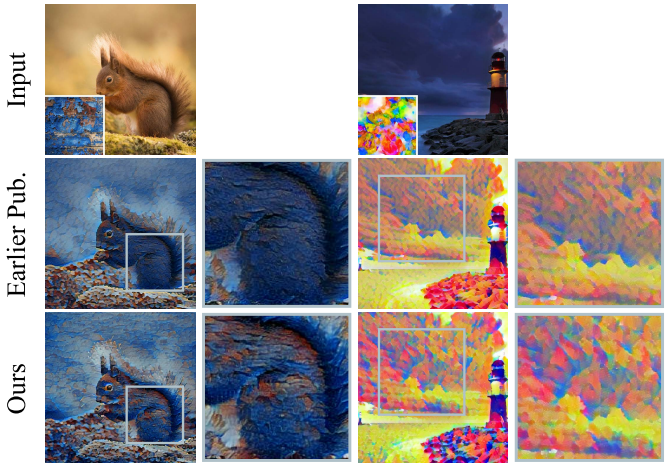


Fig. 15. Qualitative comparison against our earlier publication [12].

of these two techniques can improve temporal consistency without degrading the stylization effects. However, training with the temporal regularization can weaken the strokes and make the color duller. To alleviate this problem, we set a small temporal loss weight so that the stylization effects are still pleasing.

D. Comparison Against the Earlier Publication

Compared with our earlier publication [12], although the quantitative improvement in temporal consistency is limited as shown in Table. VI, the improvement in the stylization effect is obvious. As shown in Fig. 15, with the proposed relaxed style loss, the color becomes richer, the details are clearer, and the texture of strokes is closer to the style reference. This demonstrates the effectiveness of our new designs.

E. Effectiveness of Compound Regularization

In this section, we compare the proposed compound regularization with other regularizations.

1) Discussion: Most existing models are trained on real videos with temporal loss. However, due to the inaccuracy of optical flows, or color/illumination variations, *etc.*, the training data usually does not satisfy Eq. (1). This may result in underfitting and degrade the performance.

To avoid the trouble of collecting video data and estimating optical flows, some unsupervised single-frame regularizations

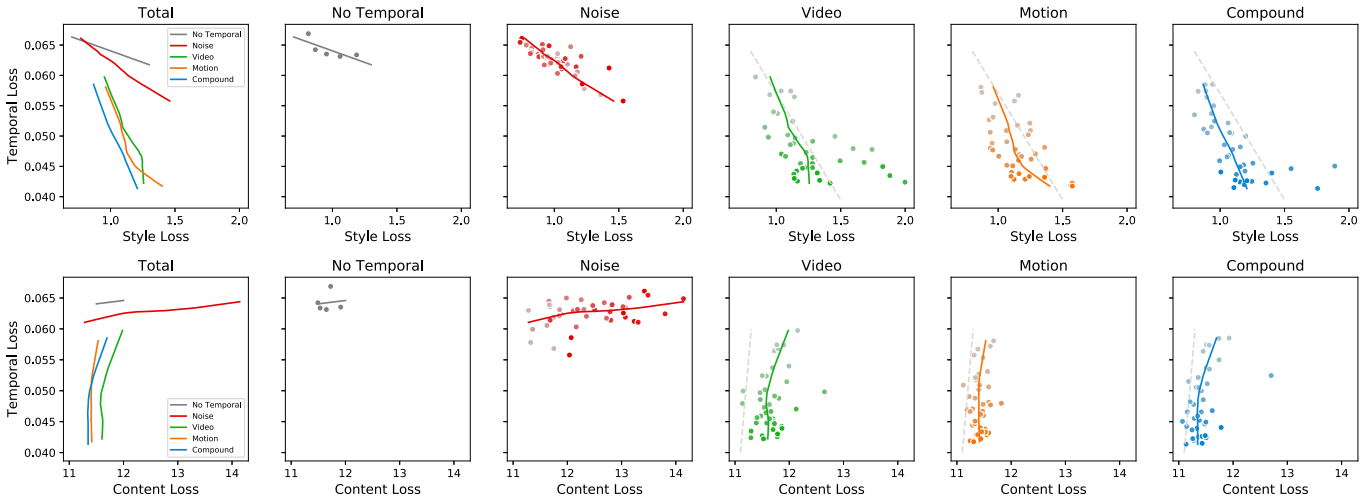


Fig. 16. Performance of the temporal consistency and stylization. Each data point represents an individual experiment. The strength of regularization is represented by different colors. A deeper color indicates a higher temporal loss weight λ_t. For the convenience of comparison, we additionally draw some light gray dotted lines as reference lines. These reference lines can better show the relative position of the points on different subfigures.

TABLE V

QUANTITATIVE ABLATION STUDY OF THE PROPOSED TECHNIQUES FOR MAINTAINING TEMPORAL CONSISTENCY. THE FULL-VERSION MODEL YIELDS THE LOWEST TEMPORAL LOSS FOR ALL TEMPORAL LENGTH

Method			Temporal Loss / Interval i				
Global	Relax	$\mathcal{L}_t$	$i=1$	$i=2$	$i=4$	$i=8$	$i=16$
			0.059	0.062	0.063	0.063	0.063
✓			0.050	0.054	0.055	0.055	0.056
	✓		0.050	0.054	0.055	0.056	0.056
		✓	0.043	0.047	0.049	0.050	0.050
✓	✓		0.046	0.050	0.051	0.051	0.052
✓		✓	0.040	0.044	0.045	0.046	0.047
	✓	✓	0.038	0.042	0.044	0.046	0.047
✓	✓	✓	**0.036**	**0.040**	**0.042**	**0.043**	**0.044**

TABLE VI

QUANTITATIVE COMPARISON AGAINST OUR EARLIER PUBLICATION [12]. OUR NEW MODEL YIELDS THE LOWEST TEMPORAL LOSS FOR ALL TEMPORAL LENGTH

Method	Temporal Loss / Interval i				
	$i=1$	$i=2$	$i=4$	$i=8$	$i=16$
Earlier Pub.	**0.036**	0.041	0.044	0.045	0.047
Ours	**0.036**	**0.040**	**0.042**	**0.043**	**0.044**

are proposed. The noise stability [45] can be rewritten with our variables as

$$\mathcal{L}_{noise} = ||\mathcal{F}(X+\Delta) - \mathcal{F}(X)||. \tag{19}$$

The transformation invariance [31] can be rewritten as

$$\mathcal{L}_{motion} = ||\mathcal{F}(W(X)) - W(\mathcal{F}(X))||. \tag{20}$$

Compared with training on videos, using regularization first generates inter-frame variations then obtains adjacent frames, therefore can obtain absolutely accurate ground truth labels. Compared with $\mathcal{L}_{comp}$, both $\mathcal{L}_{noise}$ and $\mathcal{L}_{motion}$ only contain one kind of transformation. Therefore, they cannot guide the network to optimize in the most correct direction. To demonstrate this, we benchmarked the above training techniques with our baseline image style transfer network.

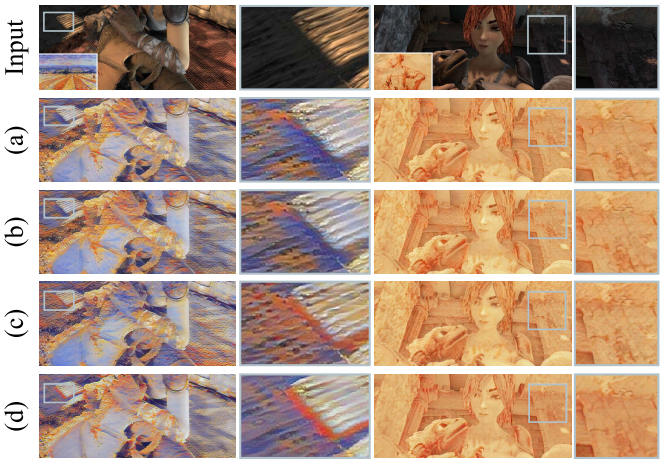


Fig. 17. Ablation study of the proposed model: (a) baseline, (b) baseline + global feature sharing, (c) baseline + relaxing style loss, (d) baseline + compound regularization.

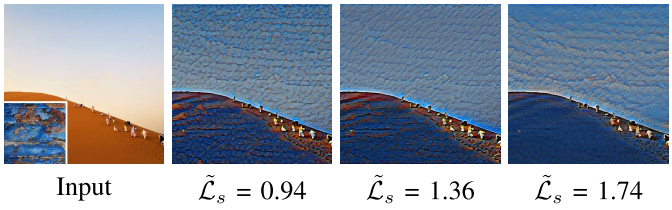


Fig. 18. Results of models with different style loss. We choose three models with similar temporal loss (0.0478 $\sim$ 0.0481) but various style loss $\tilde{\mathcal{L}}_s$. Models with higher style losses fail to well reconstruct the wooden texture.

2) Experimental Settings: For training on videos, we followed the loss function and training strategy of [9], the training data of Blind [30], and use PWC-Net [44] to estimate optical flows. $\mathcal{L}_{noise}$ and $\mathcal{L}_{motion}$ are implemented in the same

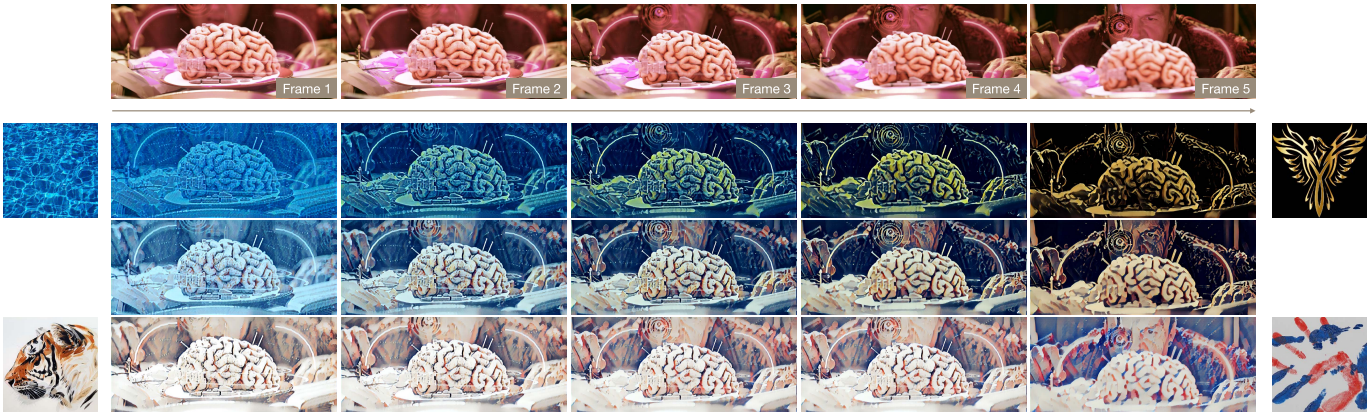


Fig. 19. Interpolation between four styles. Benefiting from single-frame processing and introducing inter-frame information in the way of global feature sharing, our model allows styles to dynamically change over time.

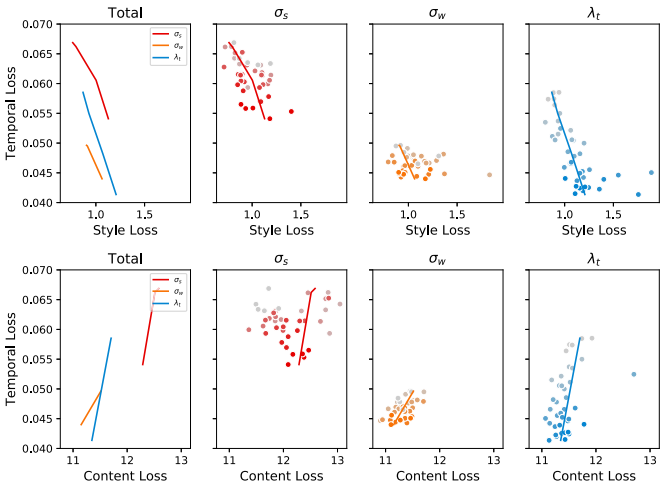


Fig. 20. Performance of compound regularization under different parameters. A deeper color indicates a higher parameter value.

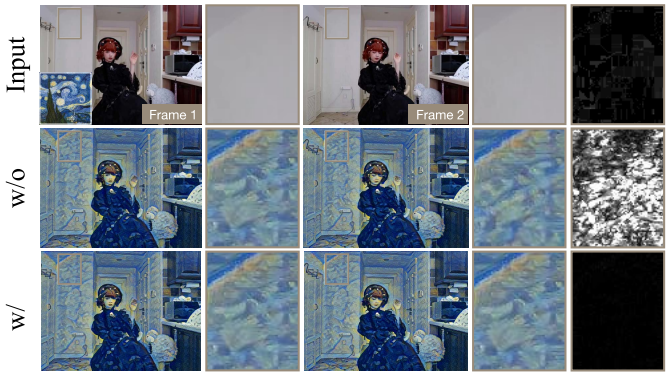


Fig. 21. Effect of with and without noise removal on degraded frames. The inter-frame residual of the magnified area is shown on the right. The input frames have blocking noise, which leads to severe texture instability in stylization results. With noise removal, the strokes are temporally consistent.

Fig. 22. With the proposed temporal regularization, the temporal consistency of CycleGAN [46] improves.

way with $\mathcal{L}_{comp}$. To reduce the impact of randomness and inappropriate weights, we conducted 5 individual trainings and selected 8 sets of parameters: $\lambda_t = k \times i, i = \{0, .., 8\}$. For $\mathcal{L}_{noise}$, we chose $k = 50$; for the other three strategies, we chose $k = 25$. To solely evaluate the performance of regularizations, we do not use relaxed style loss or global feature sharing.

Temporal smoothness is measured by temporal loss (Eq. (17)) with interval $i = 1$. Stylization effects are evaluated by style loss and content loss.

3) Results: As illustrated in Fig. 16, even without temporal loss, there is still a trade-off between temporal stability and stylization: the decrease of temporal loss can increase style loss, and content loss vice versa. This is because content images or say input frames, are themselves temporally consistent. Stylization, however, introduces variations, making it harder to preserve temporal smoothness.

$\mathcal{L}_{noise}$ decreases temporal loss. Moreover, it changes the trade-off rate between style and temporal stability, which means the same amount of style loss increase can bring more temporal loss reduction. This indicates that regularizations can lead to better network characteristics. However, there is no strong correlation between the temporal loss weight and the style loss. Moreover, the increase of regularization strength can degrade the effect of content reconstruction.

The other three strategies have better trade-off rates for both *style-temporal* and *content-temporal*. With the increase of regularization strength, temporal smoothness improves steadily. Among all strategies, $\mathcal{L}_{comp}$ performs the best for style loss. For content loss, although $\mathcal{L}_{comp}$ performs slightly worse than $\mathcal{L}_{motion}$ when the temporal loss weight is low, on the higher strength, $\mathcal{L}_{comp}$ yields the best result. Directly training on videos performs worse than $\mathcal{L}_{comp}$ and $\mathcal{L}_{motion}$. This may due to that the color/illumination of real videos is not strictly constant, and forcing networks to uniformly stylize different colors may cause conflict.

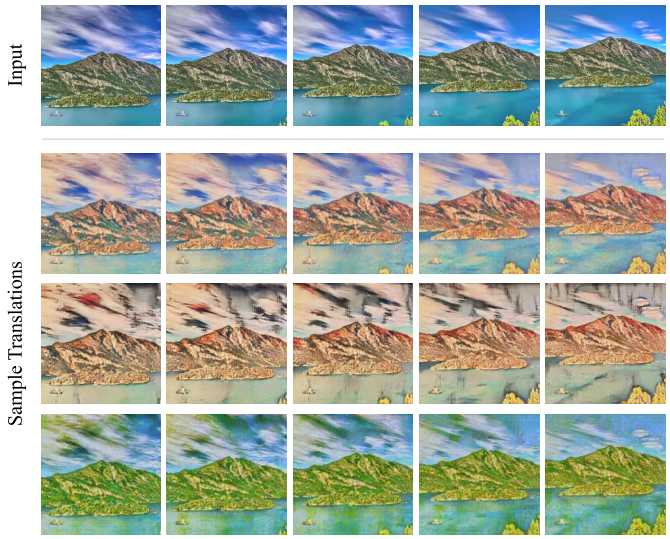


Fig. 23. Combining the proposed temporal regularization, MUNIT [47] can generate temporally smooth painting videos without hurting the generation effect and multi-model diversity.

4) Effect of Trade-Off: As shown in Fig. 18, under the same temporal loss, higher style loss can result in weaker strokes and more monotonous colors. This demonstrates the importance of a good trade-off between style loss and temporal loss.

F. Parameter Selection

In compound regularization, the magnitude of transformations is controlled by σ_s and σ_w, and the strength of regularization is controlled by λ_t. We explored the effect of these parameters. As shown in Fig. 20, σ_s, σ_w, and λ_t have similar trade-off rates for *style-temporal*. For *content-temporal*, the trade-off rate of σ_w is slightly worse than the others. Finally we set $\sigma_s = 0.01$, $\sigma_w = 8$ and $\lambda_t = 60$ for an overall good performance.

G. Application

1) Processing Degraded Videos: One underlying assumption of our model is that the input video is temporally smooth itself. However, in real applications, videos may be degraded by noises or compression artifacts. As shown in Fig. 21, the proposed model may be affected by blocking artifacts and hurt the stylization effects. A simple solution is to enhance input frames. We first compute the mean squared error of adjacent frames and set a threshold to roughly detect which object is moving. Then, for the static area, we directly copy pixels from the previous frame. As shown in Fig. 21, this simple strategy can significantly improve temporal consistency.

2) Real-Time Multiple Style Integration: Our model can integrate features to generate new styles. Moreover, benefiting from our single-frame property, styles can vary from frame to frame as shown in Fig. 19, providing users with high flexibility. This is what traditional optical-flow-based methods cannot achieve. A user interaction of real-time style adjustment is shown in the supplementary materials.

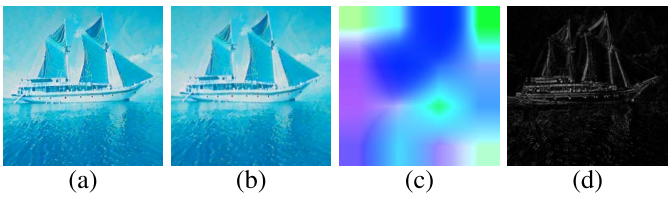


Fig. 24. Computing adversarial loss using different images. (a) is a style transfer result, (b) is warping (a) using (c), (c) is an optimized optical flow, (d) shows the residual of (a) and (b). The adversarial loss of (a) is 0.07, while the loss of (b) is 0.48.

3) Improving Other Tasks: The proposed temporal regularization can help with more than style transfer. For example, it can be used for unsupervised image-to-image translation. Fig. 22 shows the result with CycleGAN [46] on *horse2zebra*. The original temporal loss of CycleGAN is 0.090, with compound regularization, the temporal loss is 0.082. For translating photographs to paintings by MUNIT [47], temporal regularization can decrease temporal loss from 0.0479 to 0.0416. At the same time, as shown in Fig. 23, the generation visual effect and multi-model diversity is still pleasant.

VII. Conclusion and Future Work

In this article, we propose a novel video style transfer framework. To improve single-frame temporal stability, the conflict between stylization and temporal consistency is weakened through relaxing style loss. Then, we derive a new regularization term, which outperforms existing training strategies and can support various tasks. Besides, we design a sequence-level feature sharing strategy for long-term temporal consistency, and a dynamic inter-channel filter to improve the effect of stylization itself. Experimental results demonstrate the superiority of the proposed framework.

There are still some interesting issues for further investigation. A direction for future work is combining our temporal consistency techniques with GANs. We have tried to add an adversarial loss to our framework. We trained the discriminator to classify style transfer results and real artistic images. With this adversarial loss, the subjective effect is more pleasant, but the temporal consistency becomes worse. We find that like the VGG in style loss, the discriminator is also not consistent to shape variation as shown in Fig. 24. Naturally, we tried to relax the adversarial loss. However, different from the static VGG in style loss, the discriminator is dynamic during the training, therefore relaxing adversarial loss is not easy. This problem is left for future research. Improving the temporal stability of adversarial loss can be useful in many GAN-related tasks, such as video synthesis and video-to-video translation. We will try to combine our techniques with fashion synthesis [48], [49] and makeup transfer [50] in the future. We believe that our work of improving temporal consistency through relaxation and regularization can inspire researches in related fields.

Another direction for future work is considering semantics when synthesizing optical flows for compound regularization. As introduced in Sec. VI. A., we synthesize motion W in a random way, neither considering the shape of the objects nor following the laws of nature. As shown in Fig. 25,

(a) Randomly Synthesized (b) From MPI

Fig. 25. Comparison of optical flows in compound regularization and optical flows from the MPI Sintel dataset [43]. Different from randomly synthesized optical flows, real ones have object shapes and semantic meanings.

compared with our random optical flows, real motions have object shapes and semantic meanings. In this article, we chose to use random motions because that, on the one hand, it is difficult to predict possible motion patterns for each object in the static image. On the other hand, as far as we are concerned, the essence of single-frame-based video stylization is to restrict corresponding pixels to be stylized the same on different frames, which is a local or even pixel-level process. Therefore, the large-scale shape-level domain gap between random and semantic optical flows may have a limited impact on the performance. In the future, we will try to synthesize semantic optical flows and verify our assumption that shape-level gaps do not affect performance.

REFERENCES

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[2] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and superresolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2016, pp. 694–711.

[3] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "StyleBank: An explicit representation for neural image style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1897–1906.

[4] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.

[5] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4990–4998.

[6] A. J. Champandard, "Semantic style transfer and turning two-bit doodles into fine artworks," 2016, *arXiv:1603.01768*. [Online]. Available: http://arxiv.org/abs/1603.01768

[7] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Stereoscopic neural style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6654–6663.

[8] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos," in *Proc. German Conf. Pattern Recognit.*, 2016, pp. 26–36.

[9] H. Huang *et al.*, "Real-time neural style transfer for videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 783–791.

[10] D. Chen, J. Liao, L. Yuan, N. Yu, and G. Hua, "Coherent online video style transfer," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1105–1114.

[11] A. Gupta, J. Johnson, A. Alahi, and L. Fei-Fei, "Characterizing and improving stability in neural style transfer," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4067–4076.

[12] W. Wang, J. Xu, L. Zhang, Y. Wang, and J. Liu, "Consistent video style transfer via compound regularization," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12233–12240.

[13] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," in *Proc. Twenty-Sixth Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2230–2236.

[14] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 386–396.

[15] T. Qi Chen and M. Schmidt, "Fast patch-based style transfer of arbitrary style," 2016, *arXiv:1612.04337*. [Online]. Available: http://arxiv.org/abs/1612.04337

[16] L. Sheng, Z. Lin, J. Shao, and X. Wang, "Avatar-net: Multi-scale zero-shot style transfer by feature decoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8242–8250.

[17] X. Li, S. Liu, J. Kautz, and M.-H. Yang, "Learning linear transformations for fast image and video style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3809–3817.

[18] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5880–5888.

[19] Y. Yao, J. Ren, X. Xie, W. Liu, Y.-J. Liu, and J. Wang, "Attention-aware multi-stroke style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1467–1475.

[20] O. Frigo, N. Sabater, J. Delon, and P. Hellier, "Split and match: Example-based adaptive patch sampling for unsupervised style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 553–561.

[21] O. Frigo, N. Sabater, J. Delon, and P. Hellier, "Video style transfer by consistent adaptive patch sampling," *Vis. Comput.*, vol. 35, no. 3, pp. 429–443, Mar. 2019.

[22] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos and spherical images," *Int. J. Comput. Vis.*, vol. 126, no. 11, pp. 1199–1219, Nov. 2018.

[23] C. Gao, D. Gu, F. Zhang, and Y. Yu, "ReCoNet: Real-time coherent video style transfer network," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 637–653.

[24] W. Li, L. Wen, X. Bian, and S. Lyu, "Evolvement constrained adversarial learning for video style transfer," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 232–248.

[25] T. O. Aydin, N. Stefanoski, S. Croci, M. Gross, and A. Smolic, "Temporally coherent local tone mapping of HDR video," *ACM Trans. Graph.*, vol. 33, no. 6, p. 196, 2014.

[26] M. Lang, O. Wang, T. Aydin, A. Smolic, and M. Gross, "Practical temporal consistency for image-based graphics applications," *ACM Trans. Graph*, vol. 31, no. 4, p. 34, 2012.

[27] N. Bonneel, K. Sunkavalli, S. Paris, and H. Pfister, "Example-based video color grading," *ACM Trans. Graph.*, vol. 32, no. 4, p. 39, 2013.

[28] N. Bonneel, J. Tompkin, K. Sunkavalli, D. Sun, S. Paris, and H. Pfister, "Blind video temporal consistency," *ACM Trans. Graph.*, vol. 34, no. 6, p. 196, 2015.

[29] C.-H. Yao, C.-Y. Chang, and S.-Y. Chien, "Occlusion-aware video temporal consistency," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 777–785.

[30] W.-S. Lai, J.-B. Huang, O. Wang, E. Shechtman, E. Yumer, and M.-H. Yang, "Learning blind video temporal consistency," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2018, pp. 170–185.

[31] G. Eilertsen, R. K. Mantiuk, and J. Unger, "Single-frame regularization for temporally stable CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11176–11185.

[32] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[33] X. Zhang, W. Yang, Y. Hu, and J. Liu, "Dmcnn: Dual-domain multi-scale convolutional neural network for compression artifacts removal," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 390–394.

[34] J. He, C. Dong, and Y. Qiao, "Modulating image restoration with continual levels via adaptive feature modification layers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11056–11064.

[35] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "DeblurGAN-v2: Deblurring (Orders-of-Magnitude) faster and better," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8877–8886.

[36] N. Kolkin, J. Salavon, and G. Shakhnarovich, "Style transfer by relaxed optimal transport and self-similarity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10051–10060.

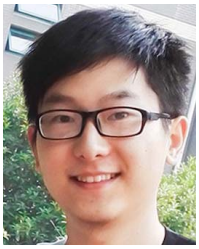
[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[38] F. Shen, S. Yan, and G. Zeng, "Neural style transfer via meta networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8061–8069.

[39] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

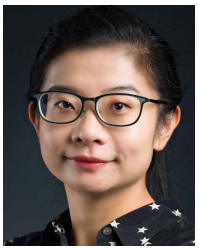
[40] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "DVC: An End-To-End deep video compression framework," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11006–11015.
[41] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering the temporal context for video question answering," *Int. J. Comput. Vis.*, vol. 124, no. 3, pp. 409–421, Sep. 2017.
[42] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D: Nonlinear Phenomena*, vol. 60, nos. 1–4, pp. 259–268, 1992.
[43] D. J. Butler, J.Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2012, pp. 611–625.
[44] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8934–8943.
[45] S. Zheng, Y. Song, T. Leung, and I. Goodfellow, "Improving the robustness of deep neural networks via stability training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4480–4488.
[46] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.
[47] X. Huang, M. Liu, S. J. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2018, pp. 179–196.
[48] C. Hsieh, C. Chen, C. Chou, H. Shuai, J. Liu, and W. Cheng, "FashionOn: Semantic-guided image-based virtual try-on with detailed human and clothing information," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 275–283.
[49] S. C. Hidayati, C. Hsu, Y. Chang, K. Hua, J. Fu, and W. Cheng, "What dress fits me best?: Fashion recommendation on the clothing style for personal body shape," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 438–446.
[50] H.-J. Chen, K.-M. Hui, S.-Y. Wang, L.-W. Tsao, H.-H. Shuai, and W.-H. Cheng, "BeautyGlow: On-demand makeup transfer framework with reversible generative network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10042–10050.

Wenjing Wang (Graduate Student Member, IEEE) received the B.S. degree in data science from Peking University, Beijing, China, in 2019, where she is currently pursuing the master's degree with the Wangxuan Institute of Computer Technology. Her current research interests include image stylization, image synthesis, and deep learning.



Shuai Yang (Member, IEEE) received the B.S. and Ph.D. degrees (Hons.) in computer science from Peking University, Beijing, China, in 2015 and 2020, respectively. He was a Visiting Scholar with Texas A&M University from September 2018 to September 2019. He was a Visiting Student with the National Institute of Informatics, Japan, from March 2017 to August 2017. He is currently a Post-Doctoral Research Fellow with the NTU AI Corporate Laboratory, Nanyang Technological University. His current research interests include image stylization and image inpainting. He received IEEE ICME 2020 Best Paper Awards and IEEE MMSP 2015 Top10% Paper Awards.

Jizheng Xu (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Shanghai Jiao Tong University, China. In 2018, he joined ByteDance Inc., as a Principal Researcher. He was a Research Manager with Microsoft Research Asia. He has authored or coauthored over 140 refereed conference papers and journal articles. He holds over 60 U.S. patents granted or pending in image and video coding. His research interests include image and visual signal representation, image/video compression and communications, computer vision, and deep learning. He is an active contributor to ISO/MPEG and ITU-T Video Coding Standards. He has over 50 technical proposals adopted by international standards, including H.264/AVC, H.264/AVC scalable extension, High-Efficiency Video Coding (HEVC), HEVC range extensions, HEVC screen content coding extensions, and versatile video coding. He chaired and co-chaired the Ad Hoc Group of exploration on wavelet video coding in MPEG and various technical ad hoc groups in JCT-VC, e.g., screen content coding, parsing robustness, and lossless coding. He was a Co-Organizer and the Co-Chair of special sessions on scalable video coding, directional transform, and high-quality video coding at various conferences. He has served as a Guest Editor for the Special Issue on Screen Content Video Coding and Applications for the IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS. He is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.



Jiaying Liu (Senior Member, IEEE) received the Ph.D. degree (Hons.) in computer science from Peking University, Beijing China, in 2010.

She was a Visiting Scholar with the University of Southern California, Los Angeles, from 2007 to 2008. She was a Visiting Researcher with Microsoft Research Asia in 2015 supported by the Star Track Young Faculties Award. She is currently an Associate Professor with the Wangxuan Institute of Computer Technology, Peking University. She has authored over 100 technical articles in refereed journals and proceedings and holds 43 granted patents. Her current research interests include multimedia signal processing, compression, and computer vision. She has served as a member of the Membership Services Committee in the IEEE Signal Processing Society, the Multimedia Systems and Applications Technical Committee (MSA TC), the Visual Signal Processing and Communications Technical Committee (VSPC TC) in the IEEE Circuits and Systems Society, and the Image, Video, and Multimedia (IVM) Technical Committee in APSIPA. She is a Senior Member of CSIG and CCF. She received IEEE ICME 2020 Best Paper Awards and IEEE MMSP 2015 Top10% Paper Awards. She has served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING and JVCI (Elsevier), the Technical Program Chair of the IEEE VCIP-2019/ACM ICMR-2021, the Publicity Chair of the IEEE ICME-2020/ICIP-2019, and the Area Chair of CVPR-2021/ECCV-2020/ICCV-2019. She was an APSIPA Distinguished Lecturer from 2016 to 2017.