Consistent Video Style Transfer via Compound Regularization

Wenjing Wang[1], **Jizheng Xu**[2], **Li Zhang**[2], **Yue Wang**[2], **Jiaying Liu**[1*]

[1] Wangxuan Institute of Computer Technology, Peking University [2] ByteDance Inc.
daooshee@pku.edu.cn xujizheng@bytedance.com lizhang.idm@bytedance.com
wangyue.v@bytedance.com liujiaying@pku.edu.cn

Figure 1: We propose a novel video style transfer framework, which can produce temporally consistent results and is highly robust to intense object movements and illumination changes. Furthermore, benefiting from the nice properties of our framework and model, we can enable features that traditional optical-flow-based methods cannot provide, such as dynamically changing styles over time. *Embedded animation best viewed in Acrobat Reader.*

Abstract

Recently, neural style transfer has drawn many attentions and significant progresses have been made, especially for image style transfer. However, flexible and consistent style transfer for videos remains a challenging problem. Existing training strategies, either using a significant amount of video data with optical flows or introducing single-frame regularizers, have limited performance on real videos. In this paper, we propose a novel interpretation of temporal consistency, based on which we analyze the drawbacks of existing training strategies; and then derive a new compound regularization. Experimental results show that the proposed regularization can better balance the spatial and temporal performance, which supports our modeling. Combining with the new cost formula, 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 other state-of-the-art style transfer methods. Our project is publicly available at: https://daooshee.github.io/CompoundVST/.

*Corresponding author. This work is partially supported by ByteDance, and partially supported by National Natural Science Foundation of China under contract No.61772043, Beijing Natural Science Foundation under contract No.L182002.

Introduction

Creating artistic imagery used to take experts hours of effort. Benefiting from the technique of style transfer, real scene images can be automatically stylized to be more visually attractive. Gatys *et al*. (2016) first proposed to use Convolutional Neural Networks (CNNs) for rendering content images, which is referred as Neural Style Transfer. Since (Gatys, Ecker, and Bethge 2016) rendered images in an iterative optimization way, which is of limited time efficiency, Johnson *et al*. (2016) proposed to do stylization in a feed-forward single-style-per-model way. Later, a variety of approaches have been proposed to further fasten the stylization process for multi-style-per-model (Chen et al. 2017b), and zero-shot style transfer (Huang and Belongie 2017). Some researches focus on extending NST for photo realistic rendering (Luan et al. 2017), doodle style transfer (Champandard 2016), and stereoscopy (Chen et al. 2018).

Applying style transfer to video is more interesting yet challenging. One difficulty is to maintain temporal consistency for video style transfer. Ruder *et al*. (2016) 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 (Huang et al. 2017; Chen et al. 2017a; Gupta et al. 2017), where picture pairs with optical flows are used to train the network with

a temporal consistency loss. However, their performance on temporal consistency is not comparable with (Ruder, Dosovitskiy, and Brox 2016).

In this paper, we try to tackle the problem of consistent video style transfer. 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 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 inter-frame relationship, which helps long-term temporal consistency. However, many models realize this by estimating optical flows, which is of limited robustness and low efficiency. We instead share global features. With feature distributions aligned among the whole sequence, networks become more robust to motions and illumination changes. One bottleneck of stylization performance lies in feature migration. Existing modules either fail to fully reconstruct style patterns or do not support end-to-end training. To address this problem, we propose to adjust inner- and inter-channel feature distributions through a dynamic filter.

Combing the above improvements, we present a video style framework. Experimental results demonstrate the superiority of the propose framework. We also show that the proposed compound regularization contributes to other vision tasks, thus can inspire researches in other domains. In summary, our contributions are threefold:

- We propose a consistent video style transfer framework with both temporally superior smoothness and visually pleasing stylization effect.
- From theoretical analysis and modeling, we derive a novel compound regularization which has superior effectiveness in guiding neural networks and can help with other computer vision video tasks.
- We develop a powerful filter to dynamically adjust inter-channel feature distributions based on both current content and style. It improves color reconstruction and enables end-to-end training.

Related Works

Image Style Transfer. Style transfer is the task of migrating styles from an artistic image to a target image. Neural Style Transfer (NST) (Gatys, Ecker, and Bethge 2016) 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 (Chen et al. 2017b) has been done to accelerate NST. 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 (Li et al. 2017a), most zero-shot methods designed feature adaptation modules. AdaIN (Huang and Belongie 2017) adjusted features through mean and variance. WCT (Li et al. 2017b) proposed multi-scale whitening and coloring transformation. Chen *et al*. (2016) migrated features by patch swapping. Avatar-Net (Sheng et al. 2018) adopted a style-swap based style decorator. Recently, Li *et al*. (2019) designed a linear transformation matrix. SANet (Park and Lee 2019) proposed to replace the patch-based mechanism with a linear module. Yao *et al*. (2019) introduced self-attention mechanism.

Existing style transfer methods fail to well balance global structures and style patterns. Moreover, most of them have no consideration of temporal consistency, therefore result in severe flickering artifacts on videos. In this paper, a novel style transfer framework is proposed, which performs better both spatially and temporally.

Video Style Transfer. Video stylization methods 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*. (2016) warped previous frames to the current time, which forms a temporal loss to guide the optimization. Based on Split and Match (Frigo et al. 2016), Frigo *et al*. (2019) also used optical flows. For further acceleration, some feed-forward networks are proposed (Gupta et al. 2017; Chen et al. 2017a; Ruder, Dosovitskiy, and Brox 2018; Gao et al. 2018; Li et al. 2018). 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, this kind of methods often neglect the spatial distribution of style patterns, 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 (Huang et al. 2017) or stable modules (Li et al. 2019). However, (Huang et al. 2017) requires an independent network for each style, while the stylization effect of (Li et al. 2019) is not satisfactory. We further explore the essence of single-frame stability maintenance and introduce sequence-level global feature sharing for better long-temporal consistency.

Some researches target at universal tasks. Lai *et al*. (2018) designed a blind post-processing neural network which supports many kinds of vision tasks. However, for style transfer, it may blur the strokes and bring out a color cast. Gabriel *et al*. (2019) 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, we further study the theoretical principle of temporal consistency; and propose an effective solution and conduct extensive experiments to demonstrate our standpoints.

[1] https://prisma-ai.com/

[2] https://deepart.io/

Temporal Consistency via Training on Single-Frame

In this section, we mathematically model temporal consistency maintenance as mapping, from which a new regularization is derived. For long-term temporal consistency, we propose a strategy of sharing global features.

Compound Regularization

Without loss of generality, we may simplify frames as vectors. Denote $X_n \in \mathbb{R}^L$ as the n-th frame of the input video. Assuming that color is constant, 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 warping X_m to X_n with the corresponding optical flow. δ is the degree of consistency. For a stable video, δ should be small enough so that human eyes are not sensitive to the flickering artifacts.

Similarly, denote $Y_n = \mathcal{F}(X_n)$ as the n-th frame of the output video. Then output temporal consistency can be written as: there exists a small number $\epsilon > 0$, such that for all n, m with $|n - m| < K$, the value of Y_n satisfies

$$||Y_n - W_{X_m \to X_n}(Y_m)|| < \epsilon. \tag{2}$$

Our target of maintaining temporal consistency 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

$$||Y' - W(Y)|| < \epsilon. \tag{3}$$

Denote $\Delta = X' - W(X)$, then we obtain

$$\begin{aligned} ||Y' - W(Y)|| &= ||\mathcal{F}(X') - W(\mathcal{F}(X))|| \\ &= ||\mathcal{F}(W(X) + \Delta) - W(\mathcal{F}(X))||, \end{aligned} \tag{4}$$

which means that improving temporal consistency is equivalent to minimizing

$$\mathcal{L}_{comp} = ||\mathcal{F}(W(X) + \Delta) - W(\mathcal{F}(X))||. \tag{5}$$

Intuitively, $\mathcal{L}_{comp}$ is a compound of two transformations: Δ represents local jitter or noise, while $W(\cdot)$ represents motions. Later we will show that $\mathcal{L}_{comp}$ can be an effective temporal consistency regularization.

Sequence-Level Global Feature Sharing

Single frame information is obviously not sufficient for stable video processing. A common way to introduce inter-frame correlation is warping frames with optical flows in the inference phase (Ruder, Dosovitskiy, and Brox 2016). However, these methods highly rely on the accuracy of optical flows and fail to handle long-term temporal consistency.

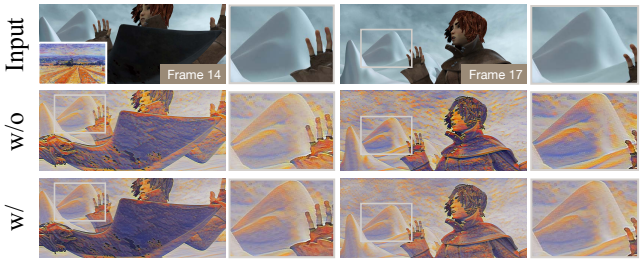


Figure 2: Ablation study of global feature sharing. With this strategy, stylized patterns of the snow mountain maintain the same appearance on frame 14 and 17.

We notice that many style transfer methods use global distributions to characterize styles, such as feature-level mean and variance in AdaIN (Huang and Belongie 2017). 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 on 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 average values are used.

As shown in Fig. 2, without sequence-level global feature sharing, stylized patterns are of different appearances on different frames, which creates flicker artifacts.

Consistent Video Style Transfer

Combining the above techniques for temporal consistency, we propose a novel video style transfer framework.

Dynamic Inter-Channel Filter

In order to improve stylization effects and enable end-to-end training, we design a new module for image style transfer.

As Li *et al.* (Li et al. 2017a) pointed out, the essence of style transfer is to migrate feature distributions. Following this idea, AdaIN (Huang and Belongie 2017) proposed to directly align the feature-level channel-wise mean and variance. However, although for every single channel the distributions are well migrated, the correlation between different channels may be still inconsistent with that of the target style. This can lead to unsatisfactory results, such as fusing colors as shown in Fig. 4(b).

Avatar-Net (Sheng et al. 2018) tried to address this problem by matching patches. However, style patterns and semantic structures are not well correlated, which may lead to messy textures and distorted contours. Moreover, it does not support end-to-end training. Yao *et al.* (2019) improved Avatar-Net with self-attention mechanisms, but the results are still unsatisfactory as shown in Fig. 5(b). SANet (Park and Lee 2019) applied a linear module, but it may distort textures as shown in Fig. 5(c).

To solve this issue, we design a new module for inter-channel feature adjustment. As shown in Fig. 3, both input and style features are fed into the Filter Predictor module to

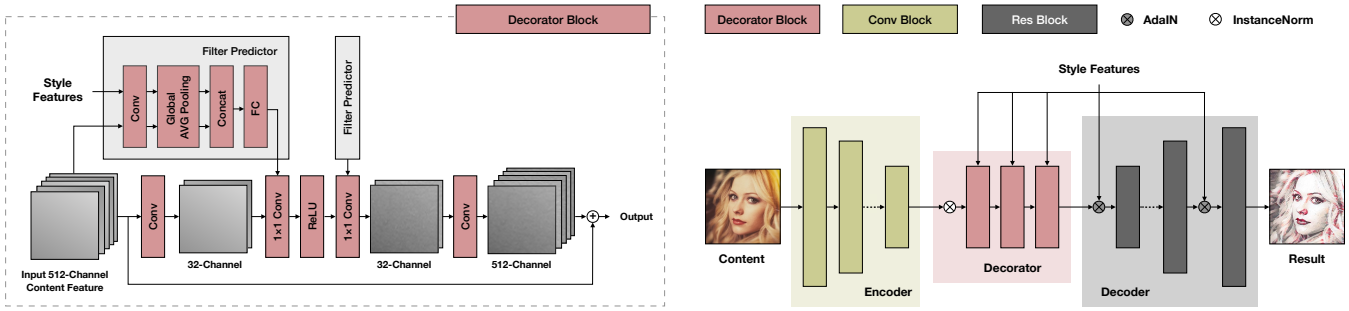
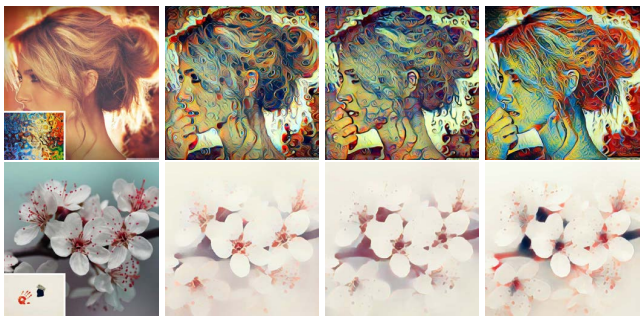
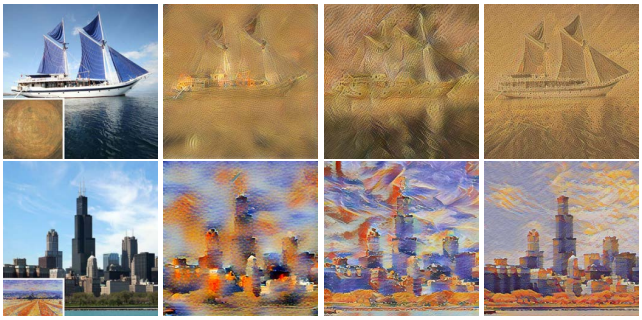


Figure 3: Left: the proposed decorator block for inter-channel feature adjustment. Both target style features and input content features are fed into a shallow sub-network Filter Predictor to predict filters. Residual learning and dimensionality reduction are used to improve the efficiency. Right: The overall architecture of the proposed encoder-decoder style transfer network.



(a) Input (b) AdaIN (c) S Filter (d) C+S Filter

Figure 4: 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.



(a) Input (b) Yao *et al.* (c) SANet (d) Ours

Figure 5: Comparison with other style transfer modules.

dynamically predict a linear combination of different channels, which is later applied to the 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. 4, 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 (Shen, Yan, and Zeng 2018), the stylization result will be unsatisfactory as shown in Fig. 4(c).

Network Architecture and Training

The architecture of the propose framework is shown in Fig. 3. We follow the hourglass encoder-decoder architecture of Avatar-Net (Sheng et al. 2018). 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. They are shared among the whole sequence in the inference phase.

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

$$\mathcal{L} = \lambda_t\mathcal{L}_t + \lambda_s\mathcal{L}_s + \lambda_c\mathcal{L}_c + \lambda_r\mathcal{L}_r + \lambda_{tv}\mathcal{L}_{tv}, \quad (6)$$

where $\mathcal{L}_t$ denotes the proposed compound temporal loss. For style loss $\mathcal{L}_s$ and content loss $\mathcal{L}_c$, we use a pre-trained VGG-19 (Simonyan and Zisserman 2014):

$$\mathcal{L}_s(S,Y) = \sum(||\text{Mean}(\Phi_l(S)) - \text{Mean}(\Phi_l(Y))||^2 + ||\text{Var}(\Phi_l(S)) - \text{Var}(\Phi_l(Y))||^2), \quad (7)$$

$$\mathcal{L}_c(C,Y) = \sum||\Phi_l(C) - \Phi_l(Y)||^2, \quad (8)$$

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. To avoid being affected by the color of content images, we desaturate content images in both training and inference phase, and introduce a color reconstruction loss:

$$\mathcal{L}_r = ||\mathcal{F}(C_{gray}, C_{color}) - C_{color}||, \quad (9)$$

where $\mathcal{F}(C_{gray}, C_{color})$ denotes colorizing gray images using the style of corresponding colorful images.

Experimental Results

Implementation Details

In compound regularization, $W(\cdot)$ is implemented by warping with a random optical flow, while Δ is a random noise

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 + Global	0.050	0.054	0.055	0.055	0.056
Baseline + Blind	0.048	0.050	0.052	0.056	0.063
Baseline + $\mathcal{L}_t$	0.041	0.045	0.048	0.049	0.050
Ours (Baseline + $\mathcal{L}_t$ + Global)	**0.036**	**0.041**	**0.044**	**0.045**	**0.047**

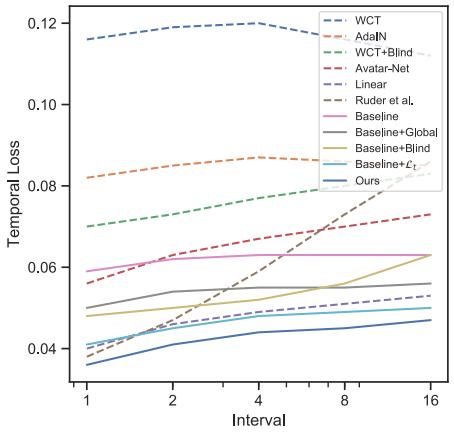


Figure 6: Quantitative evaluation of temporal consistency. For the proposed method, *Baseline* denotes the proposed image style transfer network, *Blind* denotes using Blind (Lai et al. 2018) for post-processing, $\mathcal{L}_t$ denotes training with temporal loss, and *Global* denotes using global feature sharing. Our models yields the lowest temporal loss for all temporal length.

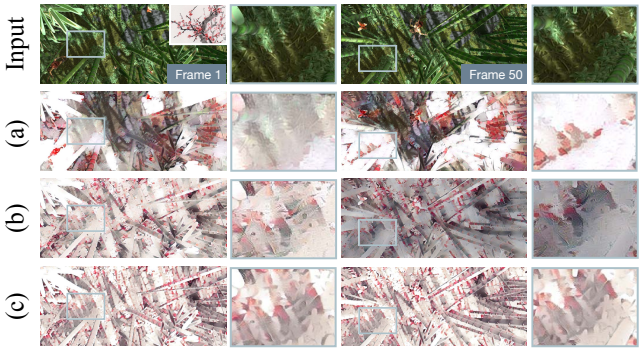


Figure 7: Comparison results on long-term temporal consistency: (a) Ruder *et al.*, (b) Baseline + Blind, (c) Ours. For (a) and (b), the bamboos are stylized differently on frame 1 and 50. Our model instead well maintains temporal consistency even with an interval of 49 frames.

with $\Delta \sim \mathcal{N}(0, \sigma^2 I)$, $\sigma^2 \sim \mathcal{U}(0.01, 0.02)$. The network is first pre-trained without $\mathcal{L}_t$ for two epochs, then fine-tuned with $\mathcal{L}_t$ for 5k iterations. More details and settings can be found in the supplementary materials.

Quantitative and Qualitative Comparisons

Temporal Consistency. The proposed method is compared with five state-of-the-art style transfer frameworks and one post-processing model for universal tasks. We evaluate both short and long-term temporal consistency:

$$\mathcal{L}_{temporal} = ||O \circ (W_{X_n \to X_{n-i}}(Y_n) - Y_{n-i}))||, \quad (10)$$

where $i \in \{1, 2, 4, 8, 16\}$ denotes frame interval, and O denotes occlusion mask. We use all the sequences of MPI Sintel dataset (Butler et al. 2012). For $i = 1$, MPI Sintel provides ground truth optical flows. For $i > 1$, we use PWC-Net (Sun et al. 2018) to estimate optical flows. Since optical flows might be inaccurate, we modify occlusion mask as

$$O' = O \cup \{||W_{X_n \to X_{n-i}}(X_n) - X_{n-i})|| > 10\}. \quad (11)$$

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

Quantitative results are shown in Fig. 6. The model proposed by Ruder *et al.* (2016) maintains good consistency for 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.* designed a blind post-processing model Blind (Lai et al. 2018). We show its result with WCT (Li et al. 2017b) (which Lai *et al.*used in the original paper) and the proposed method. Blind is also not robust to temporal length, and causes severe color bias as shown in Fig. 7(b).

For the proposed method, both global feature sharing and temporal regularization improve performance. Their combination finally yields the best result for all temporal length.

Stylization Effect. Image style transfer results are shown in Fig. 8. NST (Gatys, Ecker, and Bethge 2016) fails to balance colors. AdaIN (Huang and Belongie 2017) and WCT (Li et al. 2017b) distort content structures heavily. Avatar-Net (Sheng et al. 2018) well generates the style patterns, however, the style patterns have less correlation with the semantic structure. Linear (Li et al. 2019) introduces weird colors such as pink in the second row. Our models better balance style migration and semantic reconstruction. Compared with the baseline, the final version of our model slightly blurs the strokes, which is due to the temporal regularization. However, the result is still visually pleasing.

Video style transfer results are shown in Fig. 9. Both AdaIN, Linear, and the model proposed by Ruder *et al.* fail to reconstruct the pure blue color of the target style. WCT and Avatar-Net well synthesize the water painting stroke. However, they have high temporal errors. Compared with other methods, the proposed baseline better migrates colors and preserves semantic details. Using temporal regularization and global feature sharing, the temporal stability improves without hurting stylization effect.

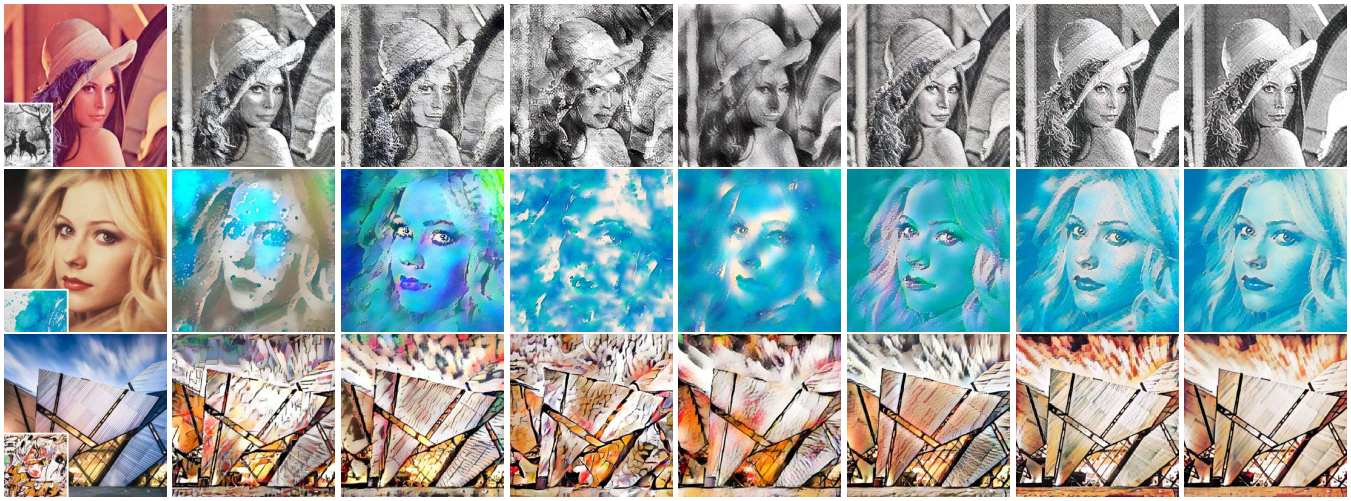


Figure 8: Comparison with state-of-the-art methods on image style transfer.

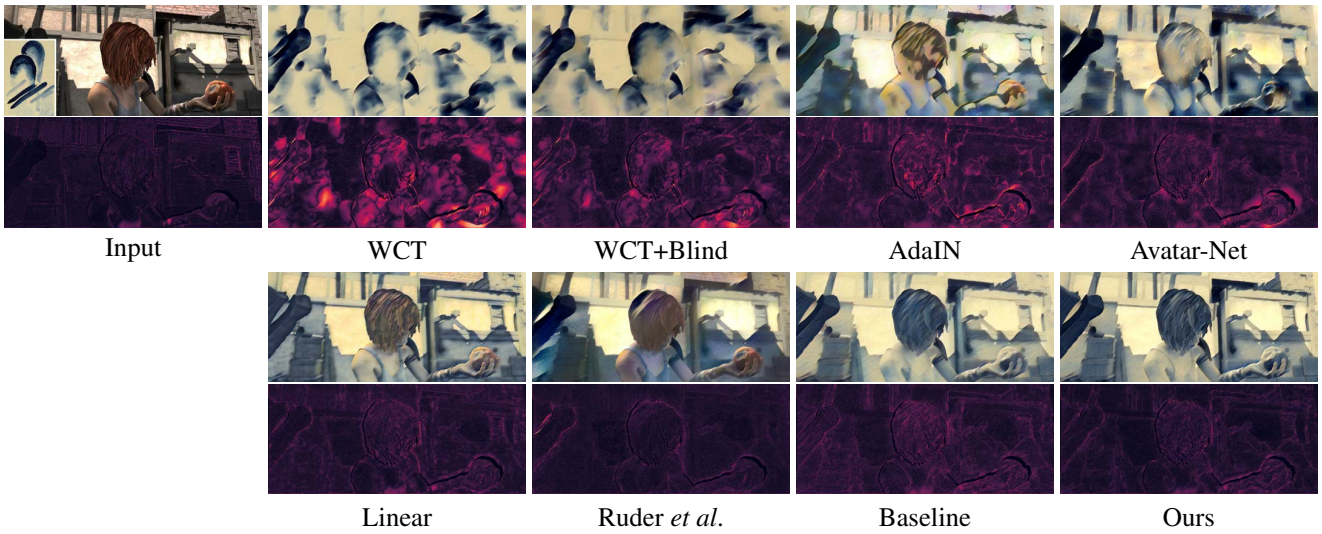


Figure 9: Comparison on video style transfer. The bottom of each row shows the temporal error heat map. Please refer to the supplementary materials for a video demonstration.

Effectiveness of Compound Regularization

Discussion. Most existing models are trained on real videos with temporal loss (Eq. (2)). However, due to, *e.g.*, the inaccuracy of optical flows, or color/illumination variations, the training data usually doesn't satisfy Eq. (1). This may result in under-fitting and degrade the performance.

To avoid the trouble of collecting video data and estimating optical flows, some unsupervised single-frame regularizations are proposed. The noise stability (Zheng et al. 2016) can be rewritten with our variables as

$$\mathcal{L}_{noise} = ||\mathcal{F}(X + \Delta) - \mathcal{F}(X)||. \quad (12)$$

The transform invariance (Eilertsen, Mantiuk, and Unger 2019) can be rewritten as

$$\mathcal{L}_{motion} = ||\mathcal{F}(W(X)) - W(\mathcal{F}(X))||. \quad (13)$$

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 benchmark the above training techniques with our baseline image style transfer network.

Experimental Settings. For training on videos, we follow the loss function and training strategy of (Huang et al. 2017), the training data of Blind (Lai et al. 2018), and use PWC-Net (Sun et al. 2018) to estimate optical flows. $\mathcal{L}_{noise}$ and $\mathcal{L}_{motion}$ are implemented in the same way with $\mathcal{L}_{comp}$. To reduce the impact of randomness and inappropriate weights, we conduct 5 individual trainings and select 8 sets of parameters. Temporal smoothness is also measured with temporal loss (Eq. (10)). Stylization effect is evaluated by style loss (Eq. (7)) and content loss (Eq. (8)).

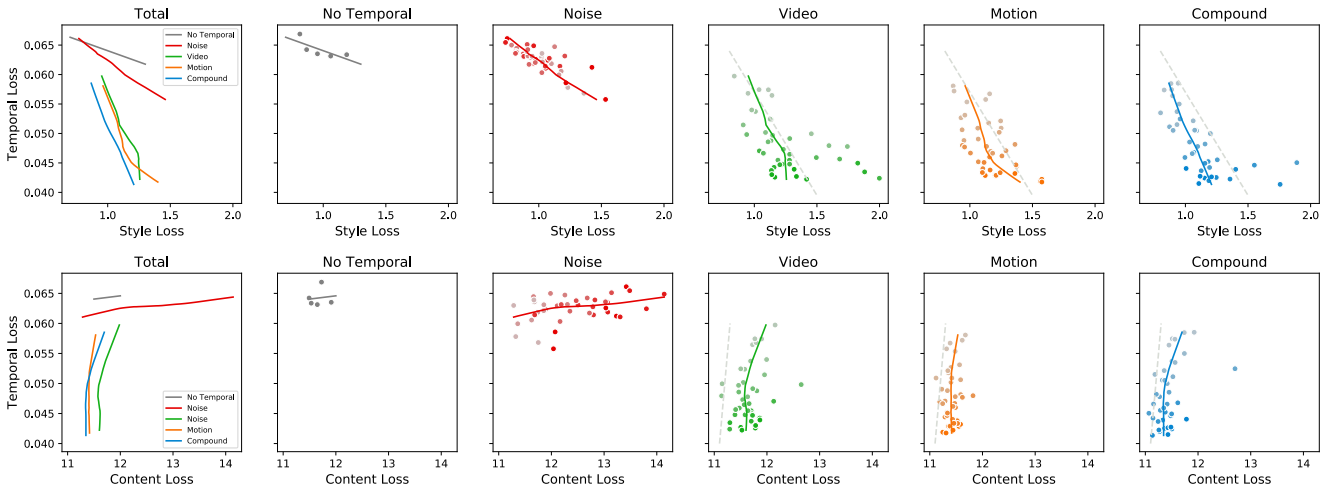


Figure 10: Performance of temporal consistency and stylization. Each data point represents an individual experiment. The strength of regularization is represented by different colors. A deeper color indicates a higher temporal loss weight. For the convenience of comparison, we additionally draw some light gray dotted lines.

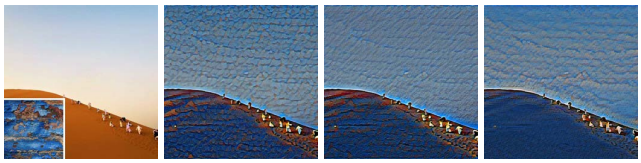


Figure 11: Results of models with different style loss. We choose three models with similar temporal loss (0.0478 $\sim$ 0.0481) but various style loss $\mathcal{L}_s$. Models with higher style loss fail to well reconstruct the wooden texture.

Figure 12: With the proposed temporal regularization, the temporal consistency of CycleGAN improves.

Results. As illustrated in Fig. 10, 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 networks characteristics. However, there is no strong correlation between temporal loss weight and style loss. Moreover, the increase of regularization strength can hurt 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 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.

Ablation Study. As shown in Fig. 11, the increase of style loss can result in weaker strokes and more monotonous colors. To alleviate this problem, we set $\lambda_t = 150$ so that with $\mathcal{L}_s = 1.067207$, the stylization effect is still pleasing.

Application

Real-Time Multiple Style Integration. Our model encodes styles into convex spaces, therefore we 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. 1, providing users with high flexibility.

Improving Other Tasks. The proposed temporal regularization can be easily applied to other vision tasks. For example, it can be used on image-to-image translation without breaking the balance of adversarial training. Fig. 12 shows the result with CycleGAN (Zhu et al. 2017) on *horse2zebra*.

Conclusion

In this paper, we propose a novel video style transfer framework. To improve single-frame temporal stability, we first derive a new regularization term, which outperforms existing training strategies and can support various tasks. Then

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. Experimental results demonstrate the superiority of the proposed framework.

References

Butler, D. J.; Wulff, J.; Stanley, G. B.; and Black, M. J. 2012. A naturalistic open source movie for optical flow evaluation. In *Proc. European Conf. Computer Vision*, Part IV, LNCS 7577, 611–625.

Champandard, A. J. 2016. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint arXiv:1603.01768*.

Chen, T. Q., and Schmidt, M. 2016. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*.

Chen, D.; Liao, J.; Yuan, L.; Yu, N.; and Hua, G. 2017a. Coherent online video style transfer. In *Proc. Int'l Conf. Computer Vision*, 1105–1114.

Chen, D.; Yuan, L.; Liao, J.; Yu, N.; and Hua, G. 2017b. Stylebank: An explicit representation for neural image style transfer. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1897–1906.

Chen, D.; Yuan, L.; Liao, J.; Yu, N.; and Hua, G. 2018. Stereoscopic neural style transfer. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 6654–6663.

Eilertsen, G.; Mantiuk, R. K.; and Unger, J. 2019. Single-frame regularization for temporally stable cnns. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Frigo, O.; Sabater, N.; Delon, J.; and Hellier, P. 2016. Split and match: Example-based adaptive patch sampling for unsupervised style transfer. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 553–561.

Frigo, O.; Sabater, N.; Delon, J.; and Hellier, P. 2019. Video style transfer by consistent adaptive patch sampling. *The Visual Computer* 35(3):429–443.

Gao, C.; Gu, D.; Zhang, F.; and Yu, Y. 2018. Reconet: Real-time coherent video style transfer network. In *Proc. Asian Conference on Computer Vision*.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2414–2423.

Gupta, A.; Johnson, J.; Alahi, A.; and Fei-Fei, L. 2017. Characterizing and improving stability in neural style transfer. In *Proc. Int'l Conf. Computer Vision*, 4067–4076.

Huang, X., and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. Int'l Conf. Computer Vision*, 1501–1510.

Huang, H.; Wang, H.; Luo, W.; Ma, L.; Jiang, W.; Zhu, X.; Li, Z.; and Liu, W. 2017. Real-time neural style transfer for videos. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 783–791.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proc. European Conf. Computer Vision*, 694–711. Springer.

Lai, W.-S.; Huang, J.-B.; Wang, O.; Shechtman, E.; Yumer, E.; and Yang, M.-H. 2018. Learning blind video temporal consistency. In *Proc. European Conf. Computer Vision*, 170–185.

Li, Y.; Wang, N.; Liu, J.; and Hou, X. 2017a. Demystifying neural style transfer. In *Int'l Joint Conf. Artificial Intelligence*, 2230–2236. AAAI Press.

Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; and Yang, M.-H. 2017b. Universal style transfer via feature transforms. In *Advances in neural information processing systems*, 386–396.

Li, W.; Wen, L.; Bian, X.; and Lyu, S. 2018. Evolvement constrained adversarial learning for video style transfer. In *Proc. Asian Conference on Computer Vision*.

Li, X.; Liu, S.; Kautz, J.; and Yang, M.-H. 2019. Learning linear transformations for fast image and video style transfer. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Luan, F.; Paris, S.; Shechtman, E.; and Bala, K. 2017. Deep photo style transfer. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 4990–4998.

Park, D. Y., and Lee, K. H. 2019. Arbitrary style transfer with style-attentional networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Ruder, M.; Dosovitskiy, A.; and Brox, T. 2016. Artistic style transfer for videos. In *Proc. German Conference on Pattern Recognition*, 26–36.

Ruder, M.; Dosovitskiy, A.; and Brox, T. 2018. Artistic style transfer for videos and spherical images. *Int'l Journal of Computer Vision* 126(11):1199–1219.

Shen, F.; Yan, S.; and Zeng, G. 2018. Neural style transfer via meta networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 8061–8069.

Sheng, L.; Lin, Z.; Shao, J.; and Wang, X. 2018. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 8242–8250.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2018. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 8934–8943.

Yao, Y.; Ren, J.; Xie, X.; Liu, W.; Liu, Y.-J.; and Wang, J. 2019. Attention-aware multi-stroke style transfer. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Zheng, S.; Song, Y.; Leung, T.; and Goodfellow, I. 2016. Improving the robustness of deep neural networks via stability training. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 4480–4488.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Proc. Int'l Conf. Computer Vision*.