Optical Flow Distillation: Towards Efficient and Stable Video Style Transfer

Xinghao Chen[1⋆][0000−0002−2102−8235], Yiman Zhang[1⋆][0000−0003−4494−4196], Yunhe Wang[1][0000−0002−0142−509X], Han Shu[1], Chunjing Xu[1], and Chang Xu[2][0000−0002−4756−0609]

[1] Noah's Ark Lab, Huawei Technologies
[2] School of Computer Science, Faculty of Engineering, University of Sydney
{xinghao.chen,zhangyiman1,yunhe.wang,han.shu,xuchunjing}@huawei.com, c.xu@sydney.edu.au

Abstract. Video style transfer techniques inspire many exciting applications on mobile devices. However, their efficiency and stability are still far from satisfactory. To boost the transfer stability across frames, optical flow is widely adopted, despite its high computational complexity, *e.g.* occupying over 97% inference time. This paper proposes to learn a lightweight video style transfer network via knowledge distillation paradigm. We adopt two teacher networks, one of which takes optical flow during inference while the other does not. The output difference between these two teacher networks highlights the improvements made by optical flow, which is then adopted to distill the target student network. Furthermore, a low-rank distillation loss is employed to stabilize the output of student network by mimicking the rank of input videos. Extensive experiments demonstrate that our student network without an optical flow module is still able to generate stable video and runs much faster than the teacher network.

Keywords: Knowledge Distillation; Optical Flow; Video Style Transfer

1 Introduction

Artistic style transfer aims to transform the artistic style of a given painting to an image and has attracted tremendous interests since the seminal work of Gatys *et al.* [10]. Plenty of works have been dedicated to improving the performance of single image style transfer from different perspectives [30, 22, 24, 29, 25, 3, 7]. Meanwhile, there is growing attention for video style transfer [19, 8, 33, 34, 9, 39] due to its wider application scenarios, *e.g.*, movie synthesis and mobile applications. Compared with single image style transfer, stylizing a video is a much more challenging task. The key problem is the flickering phenomenon of the stylized videos. Due to the motion of objects and the changing of light in the video *etc.*, transferring the videos frame-by-frame independently causes the temporal inconsistency between consecutive stylized frames.

⋆ Equal Contribution.

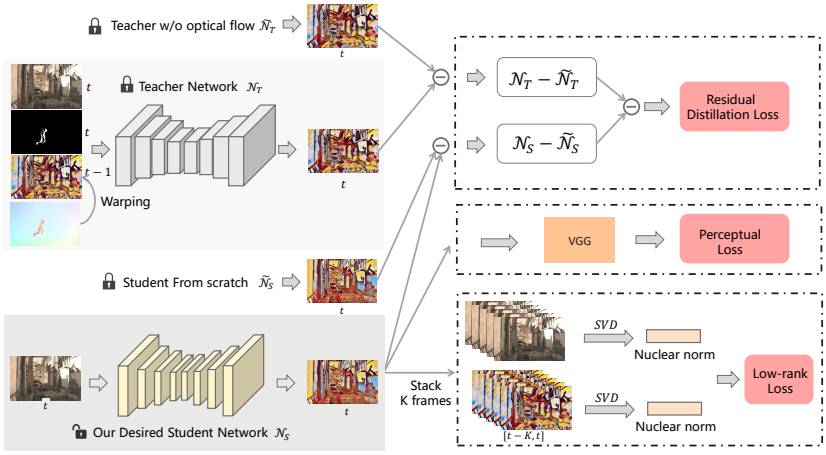

Fig. 1. The diagram of our proposed method. Our optical flow distillation method transfers knowledge from a stable video style transfer network $\mathcal{N}_T$ with optical flow to a lightweight network $\mathcal{N}_S$ that does not require optical flow during inference. The *residual distillation loss* encourages the student network to learn knowledge of the difference between the teacher networks with and without optical flow ($\mathcal{N}_T - \tilde{\mathcal{N}}_T$). Additionally, the *low rank distillation loss* is exploited to stabilize the output stylized videos of student network by mimicking the low rank of input videos. The basic *perceptual loss* is also used for style transfer.

To tackle the problem discussed above, Ruder *et al.* [33, 34] proposed a network that takes several inputs, including the current frame, occlusion mask and the previous stylized image warped by the optical flow. Despite producing smooth and coherent stylized videos, the inference speeds of these methods are relatively slow due to the calculation of optical flow on the fly. For example, the method of Ruder *et al.* [34] takes about $210ms$ to process a frame of 640×320 in the video, among which almost 97% of time is related to the time-consuming calculation of optical flow and warping operation.

Many methods have been proposed to alleviate the burden of on-the-fly optical flow computation and have achieved real-time speed for video style transfer [19, 13, 8]. These methods utilize temporal consistent loss guided by precomputed optical flow in training, which encourages the model to learn smooth and coherent stylized images in consecutive video frames. These methods are faster since they only take the current frame as input and get rid of optical flow estimation in the inference stage. Despite the fast speed, they still have less stable results when compared with methods that adopt optical flow during inference phase, as also indicated in some prior literature [8, 19]. In addition, there are a number of model compression and speeding-up methods, *e.g.*, pruning [15, 28, 37], quantization [45, 27, 36], distillation [32, 44, 26] and neural architecture search [12, 42]. Most of existing methods are explored for single image process-

ing or recognition models. An efficient algorithm for learning efficient and stable video style transfer networks is urgently required.

In this paper, we present a novel knowledge distillation method to achieve a better trade-off between inference speed and stability of the stylized videos. The framework of our proposed method is depicted in Fig. 1. We choose a stable video style transfer network including an optical flow module in the inference phase as the teacher network, and a lightweight network only consumes current frame as the desired student network. We propose the residual distillation loss to encourage the student network to learn the residual between output stylized videos produced by teacher network with and without optical flow in inference. Moreover, motivated by the fact that consistent frames have the properties of low rank, we add additional low rank loss so that student network produces coherent stylized videos that have similar low rank with input videos. The inference speed of the student network is significantly faster than that of the teacher network after removing the optical flow module.

We then carefully design the evaluation experiments on benchmarks. The results illustrate that videos generated by student network learned using the proposed optical flow distillation paradigm have similar visualization quality to that of the teacher model but obviously lower computational costs, which can be further launched in real-time on mobile devices.

The rest of the paper is organized as follows: Section 2 investigates the state-of-the-art neural style transfer methods and knowledge distillation approaches. Section 3 elaborates the proposed knowledge distillation for real-time video style transfer. In Section 4, we provide extensive experiments to compare with state-of-the-art methods and perform ablation studies. Section 5 provides a brief conclusion of this paper and discusses future work.

2 Related Work

In this section, we briefly review the related work about neural style transfer and knowledge distillation.

2.1 Video Style Transfer

Neural style transfer is one of the most popular research hotspots in recent years. Gatys *et al.* [10, 11] used CNN to iteratively reconstruct a stylized image by minimizing the difference between the target image, the content image and the style image in high-level features. These methods solve the optimization by backward propagation and are computationally expensive. To make the inference more efficient, Johnson *et al.* [22] proposed a feed-forward network to stylize images, which replaces the iterative process of optimizing pictures with the optimization of CNNs via training.

Video style transfer is attracting more and more research interests. Researchers tried to utilize the inter-frame temporal relation to improve the visual stability of stylized videos, specifically motion estimation based on optical

flow. Ruder *et al.* [33] initialized the optimization of the current frame with stylized output of the previous frame and proposed temporal loss which uses optical flow to maintain inter-frame consistency. This image based optimization algorithm outputs a very stable video but costs about 3 minutes to process a frame even with precomputed optical flow. Therefore, fast video style transfer is mainly based on model optimization. Ruder *et al.* [34] proposed a framework to use optical flow both in the training stage and in the inference stage to improve temporal consistency of output stylized videos. This framework contains two networks. The first network obtains the first frame of the video as input and outputs the stylized result. The second network obtains three inputs, including the current frame, the previous stylized frame warped by the optical flow and the mask which indicates motion boundaries and outputs the stylized result of current frame. Similarly, the architecture in [2] utilized optical flow both in training stage and inference stage. All these methods [34, 2] got stable stylized video but can not be used for real-time video style transfer. To address this problem, another family of video style transfer methods [8, 19] utilize optical flow only in the training stage thus speed up the inference. These methods utilize similar temporal loss to train the feed-forward transform network to improve the temporal consistency of output videos. They get rid of computing optical flows on the fly but produce less stable stylized results than those networks that adopt optical flows during inference stage [34, 19]. Our method aims to mitigate the gap between optical flow based and optical flow free methods for video style transfer via knowledge distillation.

2.2 Knowledge Distillation

Knowledge distillation is a technique that leverages intrinsic information of teacher network to train a smaller one, which is first pioneered by Hinton *et al.* [18]. Since then many algorithms have been proposed to improve knowledge distillation [20, 43, 6, 32, 16]. Wang *et al.* [40] exploited generative adversarial network to encourage the student network to learn similar feature distribution with teacher network. Zagoruyko and Komodakis [44] proposed to utilize spatial attention for distilling intermediate latent features of the network. Heo *et al.* [17] proposed a novel activation transfer loss to distill knowledge of activation boundaries from the teacher network. Chen *et al.* [5] introduced the locality preserving loss to preserve relationships between samples in high dimensional features from teacher network and low dimensional features from student network. Knowledge distillation has also been adopted in many other applications, *e.g.*, object detection [23, 41, 4], semantic segmentation [21, 26, 14] and pose regression [35, 38]. However, distilling knowledge from a stable video style transfer to a lightweight student network is not yet explored, which is the main purpose of this paper.

3 Method

Let $\mathcal{N}_T$ and $\mathcal{N}_S$ denote the teacher network and the desired student network, respectively. The teacher network [34] utilizes optical flow in both training and

inference phase to increase the stability of video neural style transfer. Since the teacher network needs to calculate optical flow in inference phase, it is time-consuming and not suitable for real-time applications. Thus, the student network is expected to have no optical flow module. The lightweight student network gets rid of optical flow estimation in inference phase and thus runs much faster. We choose similar network architecture with ReCoNet [8] as our student network. The goal of the proposed optical flow knowledge distillation is to train the student network $\mathcal{N}_S$ with aid of teacher network $\mathcal{N}_T$, so that $\mathcal{N}_S$ obtains similar stability as $\mathcal{N}_T$ yet still runs fast since $\mathcal{N}_S$ does not compute optical flow on-the-fly during inference.

3.1 Preliminaries: Style Transfer Loss

Here we briefly revisit the perceptual loss introduced by Johnson *et al.* [22], which has been widely used for style transfer algorithms. The perceptual loss includes the content loss to encourage the output stylized image to have similar content representations with the input image, and the style loss to capture the style information. In addition, total variation regularization ($\mathcal{L}_{tv}$) is generally introduced to encourage spatial smoothness in the stylized images. Therefore, the basic style transfer perceptual loss contains three terms:

$$\mathcal{L}_{percep}(x, I_s) = \lambda_c \mathcal{L}_{content}(x, I_s) + \lambda_s \mathcal{L}_{style}(x, I_s) + \lambda_{tv} \mathcal{L}_{tv}(x, I_s), \quad (1)$$

where I_s is the given style image and x is an input image. λ_c, λ_s and λ_{tv} are hyper parameters to balance three different losses. To simplify the notations, we omit the I_s and denote the perceptual loss as $\mathcal{L}_{percep}(x)$ in the following sections.

This basic style transfer loss can be exploited to train a student network from scratch, which is denoted as $\tilde{\mathcal{N}}_S$. Since $\tilde{\mathcal{N}}_S$ is trained frame by frame, it suffers from flickering in consecutive frames and produces unstable stylized videos. Our goal is transferring the knowledge of a teacher network $\mathcal{N}_T$ to train the desired student network $\mathcal{N}_S$, so that student network produces coherent stylized videos.

3.2 Residual Distillation Loss

A straightforward way to train the desired student network $\mathcal{N}_S$ via knowledge distillation is directly using the output stylized image of teacher network to teach the student network, as shown in the following loss function:

$$\mathcal{L}_{vanilla}(x) = ||\mathcal{N}_T(x, f) - \mathcal{N}_S(x)||_2^2, \quad (2)$$

where $\mathcal{N}_T(x, f)$ and $\mathcal{N}_S(x)$ are the output stylized images for input image x of teacher network and student network respectively, f is the corresponding optical flow.

However, this strategy produces blurred stylized images, as shown in Fig. 2. The reasons behind are two folds. Firstly, it is a challenging optimization problem to force the output of student network to align with that of the teacher

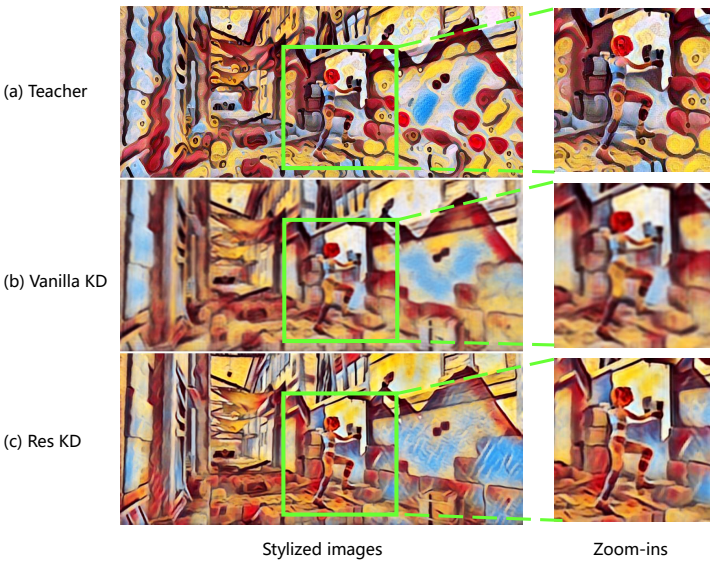

Fig. 2. (a) Output stylized frames from teacher network, (b) Results of student network with vanilla distillation loss and (c) Results of student network with residual distillation loss.

network at pixel level. More importantly, due to the difference of network architectures between student network and teacher network, the style of the output images from student network may slightly differ from those produced by teacher network. Therefore, directly distilling knowledge from the output stylized images of teacher network is hard to train a good student network.

To address the above problem, we propose the residual distillation loss for better knowledge distillation. Our goal is to train an optical flow free student network from the knowledge of an optical flow based teacher network. Therefore, the knowledge of the difference between teacher networks with and without optical flow is the key information to train a stable student network. Let $\tilde{\mathcal{N}}_T$ denotes the model similar to teacher model $\mathcal{N}_T$ but does not adopt optical flow for video style transfer and $\tilde{\mathcal{N}}_T(x)$ is the output of $\tilde{\mathcal{N}}_T$ for input image x. Obviously the output stylized video flickers since basically $\tilde{\mathcal{N}}_T$ just predicts stylized images frame by frame. The difference between the output of $\mathcal{N}_T$ and $\tilde{\mathcal{N}}_T$ is:

$$\Delta\mathcal{T}(x) = \mathcal{N}_T(x, f) - \tilde{\mathcal{N}}_T(x). \tag{3}$$

$\Delta\mathcal{T}(x)$ is attributed to additional temporal consistency that is brought by optical flows to the output stylized videos, which is the key information for student network to improve the stability.

$\tilde{\mathcal{N}}_S$ is the student network that is trained only using basic style transfer loss as in Eq. (1) and $\mathcal{N}_S$ is the stable student network we want to obtain. Thus the

improvement of $\mathcal{N}_S$ over a vanilla baseline is:

$$\Delta\mathcal{S}(x) = \mathcal{N}_S(x) - \tilde{\mathcal{N}}_S(x). \tag{4}$$

We encourage the student network to learn how to improve the stability of output videos over the baseline model by forcing $\Delta\mathcal{S}$ to imitate $\Delta\mathcal{T}$. The residual distillation loss is formulated as follow:

$$\mathcal{L}_{res}(x) = ||\Delta\mathcal{T}(x) - \Delta\mathcal{S}(x)||_2^2. \tag{5}$$

The benefits of above residual distillation loss are two folds. Firstly, $\Delta\mathcal{T}(x)$ is the key information for the teacher network to become stable and could benefit the temporal consistency of student network. What's more, $\Delta\mathcal{T}(x)$ and $\Delta\mathcal{S}(x)$ have eliminated the difference of stylized images brought by structures, thus it could alleviate the blurring effect as shown in Fig. 2.

3.3 Low-rank Distillation Loss

In the above section, Eq. (5) is proposed for distilling knowledge from teacher network for one frame. We further develop a distillation loss by exploiting the temporal property of the video.

For a stylized video that is temporally consistent and stable, the same regions that are not located at the occluded regions or motion boundaries are supposed to have similar stylized patterns and strokes. Therefore, a basic assumption for a stable stylized video is that the non-occluded regions should be low rank representations.

Suppose we have K consecutive frames $\{x_t\}_{t=1}^K$ and the corresponding optical flows $\{f_t\}_{t=1}^K$, the output stylized frames from the student network are calculated as $\{\mathcal{N}_S(x_t)\}_{t=1}^K$. We warp all stylized frames to a specific frame $\mathcal{N}_S(x_\tau)$ using the optical flows, where x_τ is often chosen as the middle frame, *i.e.*, $\tau = \lfloor K/2 \rfloor$. We denote the warped frames as $\mathcal{W}_{t\to\tau}(\mathcal{N}_S(x_t), f_t)$, where $\mathcal{W}_{t\to\tau}(\cdot)$ means warping the t^{th} frame to τ^{th} frame. $\mathcal{M}_t$ is the occlusion mask for x_t, where 0 indicates the motion regions or boundaries and 1 indicates the traceable regions. In this way, if we put all traceable regions $\mathcal{R}_t = \mathcal{M}_t \odot \mathcal{W}_{t\to\tau}(\mathcal{N}_S(x_t), f_t)$ into a matrix $\mathcal{X}$, then $\mathcal{X}$ is low rank, *i.e.*,

$$\mathcal{X} = [vec(\mathcal{R}_0), \ldots, vec(\mathcal{R}_K)]^T \in \mathbb{R}^{K\times L}, \tag{6}$$

where $L = H \times W$ is the number of pixels in the images and $vec(\cdot)$ means transforming a two dimensional image into a one-dimensional vector.

We can get the Singular Value Decomposition (SVD) of $\mathcal{X}$ by:

$$\mathcal{X} := \mathcal{U}\Sigma\mathcal{V}^T, \tag{7}$$

where $\mathcal{U} \in \mathbb{R}^{K\times K}$ and $\mathcal{V} \in \mathbb{R}^{L\times L}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{K\times L}$ is the singular value matrix. The diagonal elements in Σ are singular values $\Gamma = \{\gamma_0, \ldots, \gamma_K\}$ of the matrix $\mathcal{X}$. The rank of $\mathcal{X}$ is calculated as:

$$\text{rank}(\mathcal{X}) = \sum_i^K \mathbb{I}(\gamma_i > 0), \tag{8}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

However, the rank of a matrix is a non-differentiable function and can not directly be optimized via CNNs. Therefore, we instead adopt nuclear norm of $\mathcal{X}$, which is a convex relaxation of the rank function. The nuclear norm $||\cdot||_*$ is given by:

$$||\mathcal{X}||_* = \sum_{i}^{K} \gamma_i. \tag{9}$$

where γ_i is the i^{th} singular value of $\mathcal{X}$.

We expect that the rank of output stylized videos from student network should be similar with the stable videos. Intuitively, a straightforward way is to encourage the student network to imitate the nuclear norm of teacher network by the following low-rank distillation loss:

$$\mathcal{L}_{rank}^{T} = (||\mathcal{X}_T||_* - ||\mathcal{X}_S||_*)^2, \tag{10}$$

where $||\mathcal{X}_T||_*$ and $||\mathcal{X}_S||_*$ are nuclear norms of output stylized videos from teacher network and student network, respectively. Distilling knowledge of low rank from teacher network may help to improve the results of student network. However, the stability of output videos from the teacher network is still worse than the input videos. The stability of input video is a better teacher for the student network. To this end, we propose to distill the rank of input videos to train the student network by the following distillation loss:

$$\mathcal{L}_{rank}^{input} = (||\mathcal{X}_{input}||_* - ||\mathcal{X}_S||_*)^2, \tag{11}$$

where $||\mathcal{X}_{input}||_*$ is the nuclear norm of the input video. We utilize $\mathcal{L}_{rank}^{input}$ to train the student network and will further discuss the different low rank losses in experiments.

3.4 Optimization

Temporal consistency loss imposes constraints on consecutive output frames and is widely used in prior methods [19, 2, 8], which is formulated as follows:

$$\mathcal{L}_{temp}(x_t, x_{t-1}, f_t) = ||\mathcal{M}_t \odot (\mathcal{N}_S(x_t) - \mathcal{W}_{t-1\to t}(\mathcal{N}_S(x_{t-1}), f_t))||_2^2. \tag{12}$$

Using temporal loss improves the temporal consistency of student network and thus serves as a stronger baseline than vanilla perceptual loss. We will demonstrate the effectiveness of our proposed method on both two baselines in experiments.

To train student network with perceptual loss or residual distillation loss, the network only needs the current input frame. However, calculating low rank loss and temporal loss needs K consecutive frames. Suppose the input video segments are $x = \{x_t\}_{t=1}^{K}$ and the corresponding optical flows are $f = \{f_t\}_{t=1}^{K}$.

Algorithm 1 Optical Flow Knowledge Distillation.

Input: A given teacher network $\mathcal{N}_T$, a given vanilla teacher network $\tilde{\mathcal{N}}_T$ that is similar with $\mathcal{N}_T$ but does not use optical flow, a student network that is trained from scratch $\tilde{\mathcal{N}}_S$, a style image I_{style}, the training set $\{x^i, f^i\}_{i=1}^N$ where $x^i = \{x_t^i\}_{t=1}^K$ and $f^i = \{f_t^i\}_{t=1}^K$ are input images and optical flows, respectively.

Initialize a neural network $\mathcal{N}_S$, which does not need to compute optical flows during inference.
repeat
  Randomly select a batch of training data $\{x^i, f^i\}_{i=1}^m$.
  Residual Distillation Loss
  Employ teacher network $\mathcal{N}_T$ and $\tilde{\mathcal{N}}_T$ on the mini-batch and calculate Eq. (3).
    $\Delta T(x) \leftarrow \mathcal{N}_T(x, f) - \tilde{\mathcal{N}}_T(x)$
  Employ student network $\mathcal{N}_S$ and $\tilde{\mathcal{N}}_S$ on the mini-batch and calculate Eq. (4).
    $\Delta S(x) \leftarrow \mathcal{N}_S(x) - \tilde{\mathcal{N}}_S(x)$
  Calculate the loss function $\mathcal{L}_{res}$ according to Eq. (5).
  Baseline Loss
  Calculate perceptual loss $\mathcal{L}_{percrp}$ according to Eq. (1).
  Calculate the temporal loss function $\mathcal{L}_{temp}$ according to Eq. (12).
  Low-Rank Distillation Loss
  Calculate the nuclear norm $||\mathcal{X}_{input}||_*$ of input videos.
  Calculate the nuclear norm $||\mathcal{X}_S||_*$ of output videos from student network $\mathcal{N}_S$.
  Calculate low rank loss $\mathcal{L}_{rank}^{input}$ according to Eq. (11).
  Total Loss and Back Propagation
  Calculate the total loss function $\mathcal{L}_{total}$ according to Eq. (13).
  Update weights in $\mathcal{N}_S$ using gradient descent;
until convergence

Output: The student network $\mathcal{N}_S$.

The desired student network can be optimized using the total distillation loss as follow:

$$\begin{aligned}\mathcal{L}_{total}(x, f) = &\sum_{t=1}^{K}(\mathcal{L}_{percep}(x_t) + \lambda_{res}\mathcal{L}_{res}(x_t)) \\ &+ \sum_{t=2}^{K}\lambda_{temp}\mathcal{L}_{temp}(x_t, x_{t-1}, f_t) + \lambda_{rank}\mathcal{L}_{rank}^{input}(x, f),\end{aligned} \tag{13}$$

where λ_{res}, λ_{temp} and λ_{rank} are hyper parameters to balance different terms of losses.

The $\mathcal{L}_{percep}$ and $\mathcal{L}_{temp}$ in Eq. (13) are used as our baseline for training the student network from scratch. We will explore the baseline with and without the temporal loss. The other two terms are knowledge distillation losses proposed in this paper and we will demonstrate in experiments that these distillation losses improve the stability of output stylized videos.

4 Experiments

In this section, we will demonstrate the effectiveness of our proposed knowledge distillation for lightweight video style transfer network. In addition, we will provide extensive ablation experiments to discuss the impact of different components in our proposed method.

4.1 Experimental Settings

We use the Hollywood2 video scene dataset [31] as the training data and evaluate our method on MPI Sintel dataset [1]. We follow the same data preprocessing methods as in [34] and randomly sample 2000 tuples consisting of 5 consecutive frames from Hollywood2 dataset. MPI Sintel dataset provides ground truth optical flow and occlusion masks, which is widely used for the task of optical flow estimation and is also adopted to evaluate the temporal consistency of video style transfer. Following prior work [34, 8, 19], we evaluate our method on five videos in the MPI Sintel dataset.

During training all frames are downscaled to 640×360 and the input size for evaluation is 1024×436. We train the student network with learning rate of 10^{-3} and a batch size of 1 using Adam optimizer for 10 epochs. The learning rate is decayed by 1.2 in every 500 iterations. The hyper-parameters in Eq. (13) are set to be $K = 5$, $\lambda_{res} = 4 \times 10^8$, $\lambda_{temp} = 1 \times 10^6$ and $\lambda_{rank} = 1 \times 10^2$.

Following the quantitative evaluation metric in prior methods [8, 19, 2], we utilize e_{stab} to evaluate the temporal consistency of the output stylized videos. e_{stab} is the square root of temporal errors between consecutive frames for the traceable regions of the videos:

$$e_{stab} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\frac{1}{D}||\mathcal{M}_t \odot (y_t - \mathcal{W}(y_{t-1}))||_2}, \tag{14}$$

where N is the number of frames in the testing video, y_t and y_{t-1} are output stylized images for frame t and $t-1$ respectively, D is the number of pixels of output stylized image.

4.2 Experimental Resutls

Quantitative Analysis. We compare our proposed methods with other video style transfer networks with style *Candy* on five scenes from MPI Sintel Dataset. The temporal error e_{stab} is calculated as Eq. (14) and is used to indicate the temporal consistency of output stylized videos. Output videos with smaller values of e_{stab} are more stable. All inference speeds are evaluated on NVIDIA Tesla P100 GPU with the input size of 640×320.

As shown in Table 1, our proposed method consistently outperforms student baselines that are trained from scratch. For example, when we choose student

Table 1. Comparisons of different methods for temporal error e_{stab} and speed (*FPS*) with style *Candy* on five scenes from *MPI Sintel* Dataset. †Numbers are quoted from [8].

Models	*Alley_2*	*Ambush_5*	*Bandage_2*	*Market_6*	*Temple_2*	*Sum*	*FPS*
Teacher [34]	0.0560	0.0751	0.0489	0.0956	0.0679	0.3435	4.67
Student [8]							
- From scratch	0.0746	0.0887	0.0575	0.0997	0.0815	0.4019	**183**
- Ours	**0.0524**	**0.0676**	**0.0445**	**0.0779**	**0.0627**	**0.3050**	**183**
Student [8] w/ $\mathcal{L}_{temp}$							
- From scratch	0.0701	0.0844	0.0535	0.0948	0.0758	0.3787	**183**
- Ours	**0.0506**	**0.0643**	**0.0423**	**0.0770**	**0.0596**	**0.2938**	**183**
Chen *et al.* [2]†	0.0934	0.1352	0.0715	0.103	0.1094	0.5125	17.5
Ruder *et al.* [33]†	0.0252	0.0512	0.0195	0.0407	0.0361	0.1727	0.62

Table 2. Temporal error e_{stab} with style *Candy* for networks with different distillation losses.

Models	*Alley_2*	*Ambush_5*	*Bandage_2*	*Market_6*	*Temple_2*	*Sum*
Student	0.0746	0.0887	0.0575	0.0997	0.0815	0.4019
+ $\mathcal{L}_{res}$	0.0606	0.0764	0.0493	0.0870	0.0691	0.3424
+ $\mathcal{L}_{rank}^{input}$	0.0716	0.0862	0.0554	0.0950	0.0773	0.3855
+ $\mathcal{L}_{res}$ + $\mathcal{L}_{rank}^{input}$	**0.0524**	**0.0676**	**0.0445**	**0.0779**	**0.0627**	**0.3050**
+ $\mathcal{L}_{res}$ + $\mathcal{L}_{rank}^{T}$	0.0601	0.0755	0.0488	0.0882	0.0689	0.3415
Student w/ $\mathcal{L}_{temp}$	0.0701	0.0844	0.0535	0.0948	0.0758	0.3787
+ $\mathcal{L}_{res}$	0.0574	0.0729	0.0483	0.0870	0.0661	0.3317
+ $\mathcal{L}_{res}$ + $\mathcal{L}_{rank}^{input}$	**0.0506**	**0.0643**	**0.0423**	**0.0770**	**0.0596**	**0.2938**

network with only perceptual loss as baseline, training it with our proposed distillation losses improves e_{stab} from 0.4019 to 0.3050. To further investigate the effectiveness of our proposed method, we then switch to a strong baseline, *i.e.*, using perceptual loss and temporal loss to train the student network. Training it from scratch using additional temporal loss gets more stable results. Nevertheless, our proposed method outperforms the vanilla student by a 22.4% improvement for e_{stab}. When compared with the teacher network [34], our method achieves similar or better temporal consistency and runs much faster, since our method gets rid of time-consuming optical flow calculation during inference stage.

We further compare our proposed methods with state-of-the-art video style transfer methods [2, 33]. Table 1 shows that our method outperforms Chen *et al.* [2] for both temporal error e_{stab} and running speed. Ruder *et al.* [33] obtained smaller temporal error than our methods. However, it runs orders of magnitudes slower than our network.

Qualitative Analysis. Qualitative results of different methods are shown in Fig. 4. The first row shows two consecutive frames from the Alley_2 scene of MPI Sintel Dataset. In this video, the viewpoint of the camera is continually changing. Meanwhile, only the person in the video moves and background regions remain unchanged. Fig. 4 (b) shows the temporal consistency error of the

Fig. 3. Training the student network with (a) vanilla distillation loss $\mathcal{L}_{vanilla}$, (b) residual distillation loss $\mathcal{L}_{res}$ and (c) residual distillation loss and perceptual loss $\mathcal{L}_{percep} + \mathcal{L}_{res}$.

baseline student network, *i.e.*, which is trained from scratch with perceptual loss. It shows that the student baseline produces less temporally consistent stylized frames. The results of our method are shown in Fig. 4 (c) and achieve better temporal consistency. A stronger student baseline, *i.e.*, trained with additional temporal loss, performs slightly better than the counterpart without temporal loss. Nevertheless, our method still obtains higher temporal consistency, which demonstrates the effectiveness of our proposed method.

4.3 Ablation Studies

Impacts of Distillation Loss. We first examine the impact of our proposed distillation losses, *i.e.*, residual distillation loss $\mathcal{L}_{res}$ and low rank loss $\mathcal{L}_{rank}^{input}$. As shown in Table 2, residual distillation loss reduces e_{stab} by 15%. Adding low-rank distillation loss to the baseline network improve e_{stab} from 0.4019 to 0.3855. Furthermore, utilizing low rank loss along with residual distillation loss reduces e_{stab} by 24%. For a stronger baseline student network that adopts temporal loss, residual distillation loss and low rank loss harvest consistent improvements. These experimental results demonstrate that our proposed distillation losses effectively improve the stability of output stylized videos.

Discussion of Low-rank Loss. As we discuss in Section 3.3, there are two different design choices for low-rank distillation loss, *i.e.*, distilling low-rank knowledge from teacher network ($\mathcal{L}_{rank}^{T}$) and from input videos ($\mathcal{L}_{rank}^{input}$). As shown in the fifth and sixth row in Table 2, learning low-rank information from input videos significantly outperforms the counterpart of learning from teacher network. It's not surprising since the stability of input videos is better than output videos of teacher network. Therefore, distilling low-rank information from input videos helps to improve the temporal consistency of student network.

Table 3. Temporal error e_{stab} with style *Candy* for different K frames to calculate low rank loss.

K	3	4	5
Student (Ours)	0.3158	0.3061	0.3050
Student w/ $\mathcal{L}_{temp}$ (Ours)	0.2986	0.2950	0.2938

Residual KD vs. Vanilla KD. As we have discussed in Section 3.2, it's difficult to learn directly from the output of teacher network and thus produces blurry stylized images. As shown in Fig. 3 (a) and (b), and also in Fig. 2, training student network with $\mathcal{L}_{res}$ in Eq. (5) significantly improves the quality of stylized images compared with $\mathcal{L}_{vanilla}$ in Eq. (2).

The Impact of Perceptual Loss. An intuitive question is that, can we simply train the student network with only the information of teacher network? Fig. 3 shows that even if residual distillation loss alleviates the blurring problem, it is still not satisfying without the perceptual loss. Combining residual distillation loss and perceptual loss obtains better results with sharp edges and good style patterns. It is a reasonable observation since the task of style transfer has no groundtruth labels and it is difficult to force the student network to produce aligned outputs with teacher network in pixel-level. Therefore, the perceptual loss is still critical for the task of video style transfer knowledge distillation.

Impact of K. Here we discuss the impact of hyper parameter K, *i.e.*, the number of frames to calculate low rank loss. As shown in Table 3, increasing K from 3 to 4 slightly improves the temporal consistency of the output stylized videos. Further increasing K to 5 obtains nearly saturated improvement. To balance the performance and training cost, we choose $K = 5$ in our experiments.

5 Conclusion

In this paper, we propose a novel method to distill knowledge from a stable video style transfer network with optical flow to a lightweight network that does not require optical flow during inference. In particular, we propose the residual loss to encourage student network to learn knowledge of the difference between the teacher networks with and without optical flow. Additionally, the low rank distillation loss is exploited to constrain the output stylized videos of student network to mimic the low rank of input videos, thus to further improve the stability. Extensive experiments demonstrate that our proposed method achieves pleasing and stable stylized videos in high inference speed.

Acknowledgments. We thank anonymous reviewers for their helpful comments. Chang Xu was supported by the Australian Research Council under Project DE180101438.

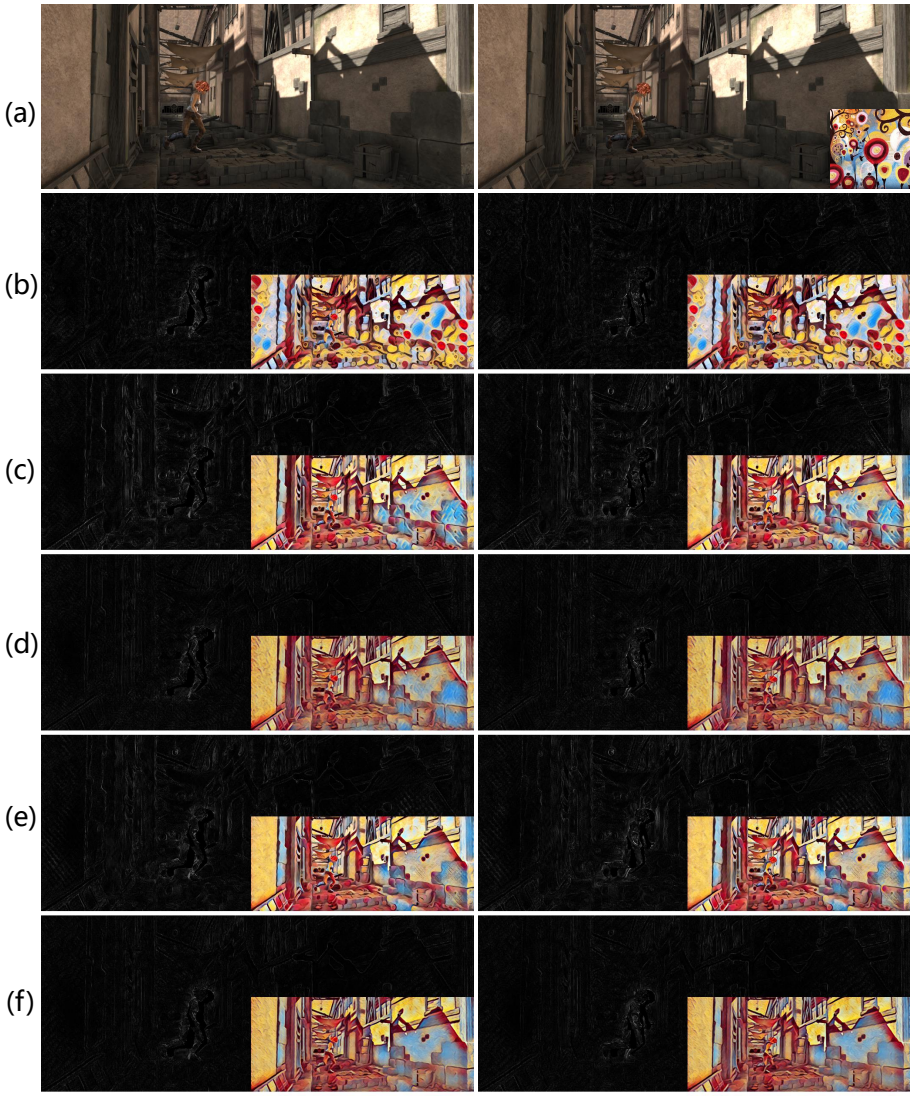

Fig. 4. Qualitative results of different methods. (a) Two consecutive input video frames from MPI Sintel Alley_2 scene. The following rows show the temporal consistency errors of (b) Teacher network, (c) Student network trained from scratch, (d) Student network by our method, (e) Student network trained from scratch with additional $\mathcal{L}_{temp}$ and (f) Student network by our method with additional $\mathcal{L}_{temp}$. The temporal errors increase as shown from black to white in gray scale. Best viewed on screen.

References

1. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: ECCV. pp. 611–625. Springer (2012)
2. Chen, D., Liao, J., Yuan, L., Yu, N., Hua, G.: Coherent online video style transfer. In: ICCV. pp. 1105–1114 (2017)
3. Chen, D., Yuan, L., Liao, J., Yu, N., Hua, G.: Stylebank: An explicit representation for neural image style transfer. In: CVPR. pp. 1897–1906 (2017)
4. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. In: NIPS. pp. 742–751 (2017)
5. Chen, H., Wang, Y., Xu, C., Xu, C., Tao, D.: Learning student networks via feature embedding. TNNLS (2020)
6. Chen, H., Wang, Y., Xu, C., Yang, Z., Liu, C., Shi, B., Xu, C., Xu, C., Tian, Q.: Data-free learning of student networks. In: ICCV (2019)
7. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. In: ICLR (2017)
8. Gao, C., Gu, D., Zhang, F., Yu, Y.: Reconet: Real-time coherent video style transfer network. In: ACCV. pp. 637–653. Springer (2018)
9. Gao, W., Li, Y., Yin, Y., Yang, M.H.: Fast video multi-style transfer. In: WACV. pp. 3222–3230 (2020)
10. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)
11. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR. pp. 2414–2423 (2016)
12. Gong, X., Chang, S., Jiang, Y., Wang, Z.: Autogan: Neural architecture search for generative adversarial networks. In: ICCV. pp. 3224–3234 (2019)
13. Gupta, A., Johnson, J., Alahi, A., Fei-Fei, L.: Characterizing and improving stability in neural style transfer. In: ICCV. pp. 4067–4076 (2017)
14. He, T., Shen, C., Tian, Z., Gong, D., Sun, C., Yan, Y.: Knowledge adaptation for efficient semantic segmentation. In: CVPR. pp. 578–587 (2019)
15. He, Y., Zhang, X., Sun, J.: Channel pruning for accelerating very deep neural networks. In: ICCV (2017)
16. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: ICCV. pp. 1921–1930 (2019)
17. Heo, B., Lee, M., Yun, S., Choi, J.Y.: Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In: AAAI. vol. 33, pp. 3779–3787 (2019)
18. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
19. Huang, H., Wang, H., Luo, W., Ma, L., Jiang, W., Zhu, X., Li, Z., Liu, W.: Real-time neural style transfer for videos. In: CVPR. pp. 783–791 (2017)
20. Huang, Z., Wang, N.: Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219 (2017)
21. Jiao, J., Wei, Y., Jie, Z., Shi, H., Lau, R.W., Huang, T.S.: Geometry-aware distillation for indoor semantic segmentation. In: CVPR. pp. 2869–2878 (2019)
22. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV. pp. 694–711. Springer (2016)
23. Li, Q., Jin, S., Yan, J.: Mimicking very efficient network for object detection. In: CVPR. pp. 6356–6364 (2017)

24. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. In: NIPS. pp. 386–396 (2017)
25. Liao, J., Yao, Y., Yuan, L., Hua, G., Kang, S.B.: Visual attribute transfer through deep image analogy. ACM Transactions on Graphics (TOG) **36**(4), 1–15 (2017)
26. Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J.: Structured knowledge distillation for semantic segmentation. In: CVPR. pp. 2604–2613 (2019)
27. Liu, Z., Wu, B., Luo, W., Yang, X., Liu, W., Cheng, K.T.: Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm. In: ECCV. pp. 722–737 (2018)
28. Liu, Z., Sun, M., Zhou, T., Huang, G., Darrell, T.: Rethinking the value of network pruning. In: ICLR (2019)
29. Lu, M., Zhao, H., Yao, A., Chen, Y., Xu, F., Zhang, L.: A closed-form solution to universal style transfer. In: ICCV (October 2019)
30. Lu, M., Zhao, H., Yao, A., Xu, F., Chen, Y., Zhang, L.: Decoder network over lightweight reconstructed feature for fast semantic style transfer. In: ICCV. pp. 2469–2477 (2017)
31. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR. pp. 2929–2936. IEEE Computer Society (2009)
32. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: ICLR (2015)
33. Ruder, M., Dosovitskiy, A., Brox, T.: Artistic style transfer for videos. In: German Conference on Pattern Recognition. pp. 26–36. Springer (2016)
34. Ruder, M., Dosovitskiy, A., Brox, T.: Artistic style transfer for videos and spherical images. IJCV **126**(11), 1199–1219 (2018)
35. Saputra, M.R.U., de Gusmao, P.P., Almalioglu, Y., Markham, A., Trigoni, N.: Distilling knowledge from a deep pose regressor network. In: ICCV (2019)
36. Shen, M., Han, K., Xu, C., Wang, Y.: Searching for accurate binary neural architectures. ICCV Neural Architects Workshop (2019)
37. Shu, H., Wang, Y., Jia, X., Han, K., Chen, H., Xu, C., Tian, Q., Xu, C.: Co-evolutionary compression for unpaired image translation. In: ICCV. pp. 3235–3244 (2019)
38. Wang, C., Kong, C., Lucey, S.: Distill knowledge from nrsfm for weakly supervised 3d pose learning. In: ICCV. pp. 743–752 (2019)
39. Wang, W., Xu, J., Zhang, L., Wang, Y., Liu, J.: Consistent video style transfer via compound regularization. In: AAAI. pp. 12233–12240 (2020)
40. Wang, Y., Xu, C., Xu, C., Tao, D.: Adversarial learning of portable student networks. In: AAAI (2018)
41. Wei, Y., Pan, X., Qin, H., Ouyang, W., Yan, J.: Quantization mimic: Towards very tiny cnn for object detection. In: ECCV. pp. 267–283 (2018)
42. Yang, Z., Wang, Y., Chen, X., Shi, B., Xu, C., Xu, C., Tian, Q., Xu, C.: Cars: Continuous evolution for efficient neural architecture search. In: CVPR. pp. 1829–1838 (2020)
43. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: CVPR. pp. 4133–4141 (2017)
44. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: ICLR (2017)
45. Zhou, S., Wu, Y., Ni, Z., Zhou, X., Wen, H., Zou, Y.: Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. arXiv preprint arXiv:1606.06160 (2016)