Temporal Denoising Mask Synthesis Network for Learning Blind Video Temporal Consistency

Yifeng Zhou[1], Xing Xu[1,*], Fumin Shen[1], Lianli Gao[1], Huimin Lu[2], and Heng Tao Shen[1]
[1]Center for Future Multimedia and School of Computer Science and Engineering,
University of Electronic Science and Technology of China, China
[2]Department of Mechanical and Control Engineering, Kyushu Institute of Technology, Japan

(a) ColorConstancy (b) Intrinsic Decomposition (c) Colorization (d) Style transfer

Figure 1: Different applications of our proposed method. It takes per-frame processed videos with serious temporal flickering as inputs (lower-left of (a) and (b)) and generates temporally stable videos while maintaining perceptual similarity to the processed frames. Our method is blind to the specific image processing algorithm applied to input videos and runs a high frame-rates. We also compare our results (upper-right) with previous state-of-the-art method [27] (lower-left of (c) and (d)). All these figures with animations are best viewed using PDF viewer of Adobe Acrobat. Kindly see more results in https://github.com/AcmMM2020supplementary/2020-ACM-MM-Supplementary.

ABSTRACT

Recently, developing temporally consistent video-based processing techniques has drawn increasing attention due to the defective extend-ability of existing image-based processing algorithms (*e.g.*, filtering, enhancement, colorization, etc). Generally, applying these image-based algorithms independently to each video frame typically leads to temporal flickering due to the global instability of these algorithms. In this paper, we consider enforcing temporal consistency in a video as a *temporal denoising* problem that removing the flickering effect in given unstable pre-processed frames. Specifically, we propose a novel model termed Temporal Denoising Mask Synthesis Network (TDMS-Net) that jointly predicts the motion mask, soft optical flow and the refining mask to synthesize the temporal consistent frames. The temporal consistency is learned from the original video and the learned temporal features are applied to reprocess the output frames that are agnostic (blind) to specific image-based processing algorithms. Experimental results on two datasets for 16 different applications demonstrate that the proposed TDMS-Net significantly outperforms two state-of-the-art blind temporal consistency approaches.

CCS CONCEPTS

• **Computing methodologies** → **Reconstruction**; **Image processing**; **Image-based rendering**.

KEYWORDS

Blind Video Processing, Temporal Consistency, Optical Flow

ACM Reference Format:
Yifeng Zhou, Xing Xu, Fumin Shen, Lianli Gao, Huimin Lu, Heng Tao Shen. 2020. Temporal Denoising Mask Synthesis Network for Learning Blind Video Temporal Consistency. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413788

1 INTRODUCTION

With the increasing popularity of deep neural networks (DNNs), various image translation methods have been proposed to guide the DNNs to produce stylized images. For example, the DNNs are taught to paint the same style as world masterpieces in the style transfer task [19, 24, 29]. The colorization methods [21, 49] can translate the gray image into a color version. However, simply applying these image-based style transfer approaches on videos frame-by-frame inevitably leads to the inferior translation result, since the *temporal consistency* among frames are ignored. Besides, due to the diverse properties of the different image-based processing algorithms such as image filtering, enhancement and colorization, applying these algorithms independently to each video frame during style transfer

*Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413788

typically leads to the phenomenon of *temporal flickering*. Therefore, to achieve temporally coherent results in videos, one primary challenge is to alleviate the effect of different image-based processing algorithms when modeling the video temporal consistency.

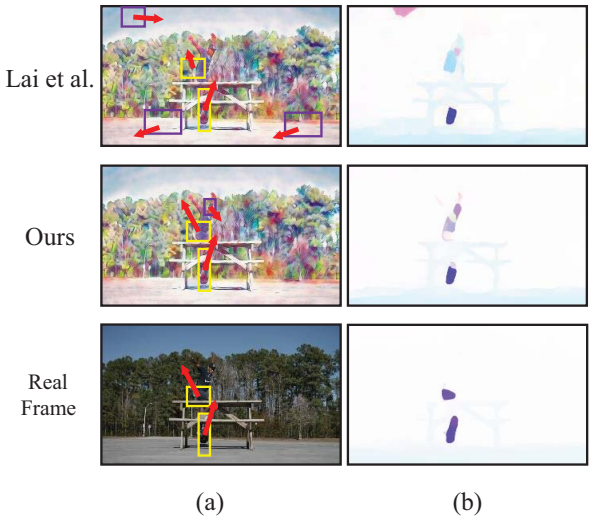


Figure 2: Our proposed model can maintain more temporal information of the background area: (a) and (b) are some example frames; (b) are the optical flow of (a). The area which is inconsistent with the optical flow of the real frame results in the problem of temporal flickering.

Recently, a few studies have been proposed to explicitly account for temporal consistency in the video, which is assumed to be *blind* to the specific image-based processing algorithms. In other words, these different algorithms are treated as black boxes that take input video frames and generate processed frames. Bonneel *et al.* [4] take the original video and the per-frame processed video as inputs and minimize the temporal warping error between consecutive frames. Yao *et al.* [45] further extend the method of Bonneel *et al.* [4] to account for occlusion by selecting a set of key-frames. The latest work of Lai *et al.* [27] formulates the problem of video temporal consistency as a learning task and minimize the short-term and long-term temporal losses between output frames in a convolutional LSTM (ConvLSTM) network. Specifically, the perceptual similarity between the output and processed frames and the spatio-temporal correlation of natural videos are fully addressed during learning.

Although the above methods have shown temporally stable results, they still have the following drawbacks: 1) Using the residual which predicted by the DNNs to synthesize the temporal consistent frames is not enough, as the residual may not well address the global flickering case. 2) The optical flow warping error will filter out some small motion information, while these are exactly the interesting features of the flickering areas. As shown in 2, the network to misled the results of the generated frame because some of the flickering areas (Purple box) are discarded by the flow warping error. 3) During the fusion period, directly adding the residual to the source frame leads the model to ignore the useful information of previous output frame. 4) For the latest work of Lai *et al.* [27], the LPIPS [50] is adopted to evaluate the perceptual distance between each output frames and input pre-processed frames. However, lower LPIPS score may not indicate better results when considering eliminating the noise in pre-processed frames.

In this paper, we proposed a novel Temporal Denoising Mask Synthesis Network (TDMS-Net) to fully overcome the shortcomings of existing approaches. Specifically, to generate a temporally consistent video sequence, we view the video temporal consistency as a *temporal denoising* task that removes the "noise" underlying the processed frames that suffering from flickering. As the framework shown in Fig. 3, the core module of TDMS contains a dual flow supervised learning module, a motion mask supervised learning module, and a refining mask unsupervised learning module. The dual flow mask which designed to teach TDMS-Net know more feature of the flicker. The motion mask which designed to take advantage of the redundant information from previous frame. Finally the refining mask is designed to synthesize a video frame with better stability and higher perceptual quality.

Our main contributions in this work is summarized as: 1) We cast the problem of video temporal consistency to the temporal denoising task and proposed a novel Temporal Denoising Mask Synthesis Network (TDMS-Net) to jointly leverage the flow, motion mask and refining mask to synthesize the high-quality and coherent video frame sequence. 2) We design four different loss terms in the proposed TDMS-Net model to fully consider the perceptual quality and temporal consistency for the resulting videos during the training process. Experiments on two widely-used dataset under 16 different applications shows the importance of all the four loss terms, and the superiority of the proposed TDMS-Net model comparing to two state-of-the-art approaches. 3) To address the incomplete evaluation of the LPIPS score proposed Lai *et al.* [27], we propose a new evaluation metric derived from the video tracking aspect, which is an useful complementary metric for the research community to assess the video temporal consistency from both the perceptual and temporal aspects.

2 RELATED WORK

Image-to-image Translation. Image to image translation is an image generation task. The goal of these tasks is to learn an image-based mapping function between the source domain and the target domain. With the development of the Generative Adversarial Networks (GANs) [12], there are emerging a lot of approaches to solve this task [23, 37, 46, 51]. In addition, some researchers also use the Variational Autoencoder (VAE) to solve this problem [26, 32]. These approaches all achieve impressive results. Our approach aims to extend image processing techniques to video. Different from the video-to-video translation, our method is the reprocessing process on the result of image-based translation methods.

Video Temporal Consistency. Simply applying the image-to-image translation method to the video frame by frame may result in a bad temporal consistency. The synthesis video is unstable because the image-based methods process the frames independently and ignore the temporal aspect. In order to teach the network to learn the relationship between two consecutive frames, many researchers add the optical flow data to their network and achieve

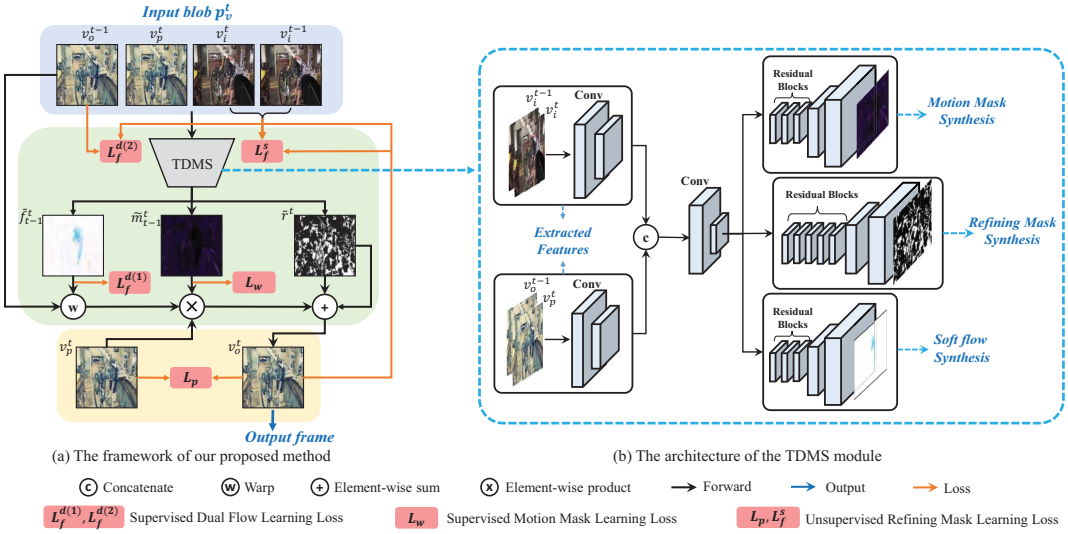


Figure 3: The overview of TDMS-Net (a) and its core module TDMS (b). Given a specific time step t, we show how TDMS-Net processes the input blob p_v^t. For the TDMS module, it contains three paralleled branches that synthesize the 1) soft flow $\tilde{f}_{t-1}^t$, 2) the motion mask $\tilde{m}_{t-1}^t$ and 3) the refining mask $\tilde{r}^t$, respectively.

good results. In the video style transfer task [8, 13, 18], the L2 loss between the warped frame and its successor or predecessor frame is regarded as the temporal loss. In their models, both the temporal loss and perceptual loss [24] are utilized in optimization. Instead of only focusing on a specific task, several researchers propose the task-independent approaches. Bonneel *et al.* [4] solve the video's flickering problem in the view of gradient-domain. Lang *et al.* [28] take use of the edge-aware filtering to improve the optical flow estimation module and create the final result. Lai *et al.* [27] embed the ConvLSTM to their model. Meanwhile, they not only compute the short-term temporal loss, but also calculate the long-time temporal loss by the frame pairs which are sampled from the output sequence. Zhang *et al.* [48] propose a ST-CLSTM structure which extracts both spatial features and temporal correlations to retain the temporal consistency. In this paper, we teach our network to learn the temporal consistency from the original videos. Then, the learned consistency is applied to the frame sequences which are processed by the image-based translation methods.

In addition, recently many researchers take use of the cycle consistency to further improve the temporal coherence of video. Wang *et al.* [38] propose a backward and forward time route to locate the aim area. Dwibedi *et al.* [10, 35, 43] encoded the video in an embedding space. Then, by learning the cycle consistency representation, the network can align the similar videos. Specifically, we are not the first to introduce dual flow to improve the model. Niklaus *et al.* [31] use the bidirectional optical flow to warp the consecutive frames. Then take both the warp frames as input to synthesize the target frame. Yang *et al.* [44] notice that only using one type of the optical flow is not enough, so they try to use the fine flow and the regularized flow to extract different temporal features and achieve an impressive result. In these approaches, the temporal consistency is learned and applied to the same domain. Differently, our method needs to apply the learned temporal feature to another domain.

Optical Flow. Optical flow is a fundamental problem in computer vision. Since Horn *et al.* [16] first presented a method for finding the optical flow pattern. A large number of flow algorithms based on the optical flow constrain [6] served as extended versions to solve this problem [5, 30, 34, 40]. With the population of convolutional neural networks, Weinzaepfel *et al.* [41] takes use of the convolutional layers and the pooling layers to improve their algorithms.

Meanwhile, many researchers try to solve this problem by directly using the end-to-end fully convolutional network [9, 22]. These methods take the frame pairs as input to predict every pixel's motion state and achieve great results in the optical flow datasets [1, 7]. However, because it is difficult to collect the optical flow datasets in the real-world, the supervised learning strategies has its bottleneck. Many researchers use the unsupervised approaches to training their networks. Yu *et al.* [47] add the brightness constancy and motion smoothness constrain to enhance their unsupervised training. Inspired by these works, we design the dual optical flow to guide our TDMS-Net to maintain the temporal coherence.

3 PROPOSED METHOD

3.1 Preliminary

Problem Formulation. Given an arbitrary video sequence $\boldsymbol{v_i} = \{v_i^1, v_i^2, \ldots, v_i^T\}$. Firstly, an image-based algorithm $F_i(\cdot)$ is used to process each frame in v_i to obtain the pre-processed video sequence $\boldsymbol{v_p} \equiv \{v_p^1, v_p^2, \ldots, v_p^T\}$. Notably, there are random flickering areas in $\boldsymbol{v_p}$, which can be viewed as the temporal noise. To generate non-flickering video frames $\boldsymbol{v_o} \equiv \{v_o^1, v_o^2, \ldots, v_o^T\}$ and to maintain the quality of the pre-processed video frames, a denoising model $D_n(\cdot)$ is utilized to reorganize the spatio-temporal features which are extracted from the input blob $\boldsymbol{p_v^t}$ (current and previous real frames v_i^t, v_i^{t-1}, generated frame v_o^{t-1} of the previous time step $t-1$; current pre-processed frame v_p^t). In summary, the above

process of generating the stable video frames can be derived as: $\boldsymbol{v_o} = D_n(F_i(\boldsymbol{v_i}))$.

Network Architecture. The framework of the proposed TDMS-Net is illustrated in Fig. 3 (a). For any video and its correspond pre-processed version, our network aims to produce consecutive frames with high temporal consistency. For each video sequence $\boldsymbol{v_i}$ and its pre-processed version $\boldsymbol{v_p}$, we sample the input blob $\boldsymbol{p_v^t}$ frame-by-frame. It is notable that the first output frame v_o^1 is assumed to be equal to v_p^1. Then, we feed the $\boldsymbol{p_v^t}$ to the temporal denoising mask synthesis (TDMS) module as elaborately shown in Fig. 3 (b).

Inspired by Xu *et al.* [42], the TDMS module firstly divides the $\boldsymbol{p_v^t}$ into two parts and then their features are extracted and down-sampled. Next, these down-sampled features are concatenated together and fed into another convolutional layer. Three sub-streams based on previous results are utilized to synthesize the soft optical flow $\tilde{f}_{t-1}^t$, the motion mask $\tilde{m}_{t-1}^t$ and the refining mask $\tilde{r}^t$. As shown in Fig. 3 (a), we first get the raw future frame by warping the v_o^{t-1} by $\tilde{f}_{t-1}^t$. Then we use the motion mask to mix the features of v_p^t and utilize the refining mask to refine the former output by eliminating the unsatisfying area.

3.2 Temporal Denoising Mask Synthesis

As illustrated in Fig. 3 (b), the TDMS can be viewed as the temporal denoising module that generates the coherent processed video sequence. The unprocessed video sequence v_i has strong consistency in the temporal dimension, while the pre-processed video sequence $\boldsymbol{v_p}$ is unstable. Therefore, the input blob $\boldsymbol{p_v^t}$ is divided into two parts: the real video part $\boldsymbol{v_{ti}} \equiv \{v_i^{t-1}, v_i^t\}$ and the processed part $\boldsymbol{v_{tp}} \equiv \{v_o^{t-1}, v_p^t\}$. Then, different filters are used to extract the temporal features and the target domain content features. While the two features tensors are down-sampled to the target size, they are concatenated together to get the mixed feature.

Supervised Dual Flow Learning. To keep the temporal coherence, the first task for the TDMS module is to predict the optical flow $\tilde{f}_{t-1}^t$. The most intuitive way is to use the FlowNet [22] to estimate the flow $\tilde{f}_{t-1}^t$, however, directly using the entire FlowNet will drastically increase the computational cost. Alternatively, we use the residual blocks [15] to achieve a simple version of the FlowNet and we consider it as the soft flow prediction. This flow generally describes the changes between the consecutive frames and it can be used to warp the last frame forward and finally improve the temporal consistency of the output frames.

In the training phase, the optical flow f_{t-1}^t between the $\boldsymbol{v_{ti}}$ is viewed as the ground truth. To guide the TDMS module learn the temporal feature, the predicted soft flow is expected to be similar to the real optical flow f_{t-1}^t. To this end, we use the L_1 distance to measure the first soft flow loss $\mathcal{L}_f^{d(1)}$:

$$\mathcal{L}_f^{d(1)} = \sum_{t=2}^{T} \|\tilde{f}_{t-1}^t - f_{t-1}^t\|_1, \tag{1}$$

where T is the number of sampled frames.

As shown in Fig. 4, based on the flow f_{t-1}^t, the naive temporal consistent loss $\mathcal{L}_f^s$ (mentioned in Section 3.2) can tell the model how to align the frames due to changing motion patterns of the cameras or objects. However, the synthesis network may add some noise that can not be suppressed by the alignment operation. If the model is blind to this noise, the model will be deceived or regard flickering frames as perfect output. For example, when the TDMS module produces a frame where the foreground objects identically comply with the temporal consistency while the background is full of unstable noise, the loss function $\mathcal{L}_f^s$ can not guarantee the synthesis frame to be perfect.

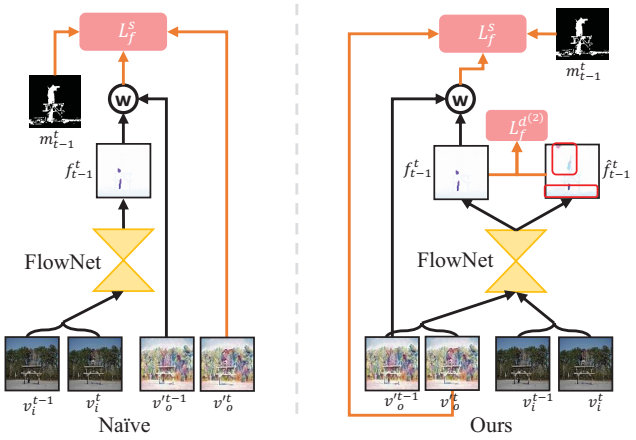


Figure 4: The naïve method which used in [18, 27, 39], only consider about one direction. We add the constrains between $\hat{f}_{t-1}^t$ and f_{t-1}^t to help our model learn more information about the temporal consistency. The Generative model may add some noise to the output frame (red rectangle), and this noise is not visible to f_{t-1}^t and m_{t-1}^t. The $v_o'^{t-1}$ and $v_o'^{t-1}$ are the results from [27].

To solve the above problem, we add an optical flow constraint in the target domain. This constraint and the above soft flow constraint are called dual flow constraint. Unlike the previous methods, we also estimate the flow $\hat{f}_{t-1}^t$ of $\boldsymbol{v_o}$. Then the second soft flow loss $\mathcal{L}_f^{d(2)}$ is defined as the L_1 distance between the f_{t-1}^t and $\hat{f}_{t-1}^t$ as

$$\mathcal{L}_f^{d(2)} = \sum_{t=2}^{T} \|f_{t-1}^t - \hat{f}_{t-1}^t\|_1. \tag{2}$$

Finally, the whole dual flow loss $\mathcal{L}_f^d$ is the summation of the two soft flow loss terms, as:

$$\mathcal{L}_f^d = \mathcal{L}_f^{d(1)} + \mathcal{L}_f^{d(2)}. \tag{3}$$

Supervised Motion Mask Learning. After receiving the soft optical flow $\tilde{f}_{t-1}^t$, the motion mask $\tilde{m}_{t-1}^t$ prediction part is designed to utilize the redundant information between the neighbor frames. The value of the motion mask varies from 0 to 1. For the mutable area, the value of $\tilde{m}_{t-1}^t$ tends to be 1, which suggests to adopt the feature of v_p^t, while for redundant area, the value of $\tilde{m}_{t-1}^t$ tends to be 0, which suggests to reuse the feature of warped v_o^{t-1}. During the training phase, the TDMS module aims to mix the warped v_o^{t-1} and v_p^t. Similar as [27, 39], using the occlusion map is an effective way to supervise our model to learn the motion mask. Then, the

motion mask loss can be formulated as:

$$\mathcal{L}_w = \sum_{t=2}^{T} \|\tilde{m}_{t-1}^t - m_{t-1}^t\|_1, \tag{4}$$

where the m_{t-1}^t is the ground truth of the occlusion map which can be calculated by $m_{t-1}^t = \exp(-\alpha\|v_i^t - F_w(v_i^{t-1}, f_{t-1}^t)\|_2)$. Empirically the α here is equal to 50.

After getting the soft optical flow and motion mask, the raw synthesis frame can be computed by:

$$v_o^t = \tilde{m}_{t-1}^t \odot F_w(v_o^{t-1}, \tilde{f}_{t-1}^t) + (1 - \tilde{m}_{t-1}^t) \odot v_p^t, \tag{5}$$

where the $\tilde{m}_{t-1}^t$ is the motion mask, the v_o^{t-1} is the last output frame, the v_p^t is the current frame of pre-processed version, and the $\tilde{f}_{t-1}^t$ is the soft optical flow. The function $F_w(v_o^{t-1}, \tilde{f}_{t-1}^t)$ denotes warping the v_o^{t-1} based on the $\tilde{f}_{t-1}^t$.

Unsupervised Refining Mask Learning. Though the previous supervised learning stages have already learned to extract the spatio-temporal features, here we also introduce the additional unsupervised learning step to refine the output of the TDMS module to produce more effective results. Considering the raw synthesis frame may be distorted due to the warped v_o^{t-1} is distinct from v_p^t and the temporal information may lose during the fusion operation, we further employ a refining mask $\tilde{r}^t$ to enable the proposed model to refine the output of the TDMS module. This refining mask is utilized to fine-tune the output by suppressing the unsatisfying area in the frames:

$$v_o^t = \tilde{m}_{t-1}^t \odot F_w(v_o^{t-1}, \tilde{f}_{t-1}^t) + (1 - \tilde{m}_{t-1}^t) \odot v_p^t + \tilde{r}^t. \tag{6}$$

Since the real coherent video is unavailable during training, the TDMS-Net is expected to distinguish the results which are more likely to have both temporal consistency and high perceptual quality via the unsupervised learning manner.

(1) *Short-term Temporal Consistency.* The short-term temporal loss $\mathcal{L}_f^s$ is defined as the warping error between the current and previous output frames. We assume the backward flow f_{t-1}^t, which was estimated by the FlowNet, as the real optical flow of the $\boldsymbol{v_{ti}}$. After using this flow to warp the previous output frame v_o^{t-1}, we can compute the L_1 loss between the current output frame v_o^t which was constructed by (6) and the warped image. In addition, we also utilize the occlusion map m_{t-1}^t to compute warping loss. Finally, the short-term temporal loss is given by:

$$\mathcal{L}_f^s = \sum_{t=2}^{T} m_{t-1}^t \odot \|v_o^t - F_w(v_o^{t-1}, f_{t-1}^t)\|_1. \tag{7}$$

(2) *Content Perceptual Consistency.* Different from the previous video transfer task, our algorithm is blind to the method used in the pre-processing phase. To produce the frame which has the same perception as that of the pre-processed frame, the content perceptual loss [24] between the output and its correspond pre-processed version need to be minimized. The content perceptual loss $\mathcal{L}_p$ is given by:

$$\mathcal{L}_p = \sum_{t=2}^{T} \sum_{l=1}^{L} \|P_l(v_o^t) - P_l(v_p^t)\|_1, \tag{8}$$

where T is the number of sampled frames, L is the number of layers and $P_l(\cdot)$ denotes the l^{th} layer's output of the feature extract network.

3.3 Objective Function

To formulate the final objective function, the perceptual content loss $\mathcal{L}_p$ in Eq. 8 is firstly used to maintain the perceptual quality. Different from the previous works [13, 18, 27] which only use the single temporal loss for training, we not only consider the flow to compute the short-term temporal loss $\mathcal{L}_f^s$ in Eq. 7, but also incorporate the dual flow loss $\mathcal{L}_f^d$ in Eq. 3. Moreover, The motion mask loss $\mathcal{L}_w$ in Eq. 4 is computed by the occlusion mask and the motion mask. Finally, the overall objective function of our TDMS-Net is to minimize the integration of the above four loss terms by:

$$\mathcal{L}_{final} = \lambda_p \mathcal{L}_p + \lambda_{fs} \mathcal{L}_f^s + \lambda_{fd} \mathcal{L}_f^d + \lambda_w \mathcal{L}_w, \tag{9}$$

where λ_p, λ_{fs}, λ_{fd} and λ_w are four hyper-parameters that control the impact of each loss term. In practice, minimizing the objective loss $\mathcal{L}_{final}$ in Eq. 9 can be performed using stochastic gradient descent (SGD) algorithm, which can compute the gradients and update the model parameters iteratively.

4 EXPERIMENT

4.1 Experimental Setup

Datasets and Pre-process. We use the same two datasets as the previous works [4, 27]. The first dataset is DAIVS-2017 [33], which contains 90 videos and is commonly used for video segmentation. Here we use 60 video sequences in this dataset to train our temporal consistency model and the remaining 30 videos are for testing. Another dataset VIDEVO is provided by Lai *et al.* [27], which contains 100 high-quality videos (80 for training and 20 for testing) crawled from Videvo.net website.

Each video in the two datasets is processed by various image-based translation algorithms depending on the specific tasks. For the style transfer methods, we use [24, 29] to pre-process the two datasets. We also use the CycleGAN [51] to translate the video sequences to other target domain (*e.g.*, the horse to zebra, the photo to painting). For other image translation methods like the colorization task, we apply the approaches proposed in [21, 49] to pre-process the videos. For the image enhancement task, we use the model of Gharbi *et al.* [11] to pre-process the videos. In addition, for the intrinsic image decomposition task, we adopt the methodology proposed by Bell *et al.* [2]. It is worth mentioning that all the above image-based translation methods output obviously flickering results which are poor to maintaining the temporal consistency.

Implementation Details. In [27], 16 pre-processed versions of the two datasets are provided. In our experiment, we use the following versions to train the TDMS-Net model: (1) the pre-processed version of WCT-antimono; (2) the pre-processed version of WCT-candy; (3) the pre-processed version of WCT-sketch; (4) the pre-processed version of DBL-expertA; (5) the pre-processed version of colorization; and (6) the pre-processed version of Intrinsic-shading. The total training set contains 840 videos. When training our TDMS-Net model, the batch size is set to 4 and each batch has 11 frames.

Table 1: Quantitative evaluation on the two metrics of temporal consistency and composited spatio-temporal consistency for our TDMS-Net and the two compared approaches.

Task	DAVIS			VIDEVO		
	Bonneel *et al.* [4]	Lai *et al.* [27]	**TDMS-Net**	Bonneel *et al.* [4]	Lai *et al.* [27]	**TDMS-Net**
WCT [19]/antimono	0.5433 / 0.5994	0.5068 / 0.6263	**0.5596** / **0.6482**	**0.6573** / 0.7122	0.5326 / 0.7495	0.6435 / **0.7832**
WCT [19]/asheville	**0.5679** / 0.6081	0.4617 / 0.6014	0.5233 / **0.6309**	**0.6314** / 0.7062	0.4743 / 0.7217	0.6151 / **0.7683**
WCT [19]/candy	**0.6294** / 0.6273	0.5083 / 0.6245	0.5618 / **0.6495**	**0.6977** / 0.7142	0.5267 / 0.7440	0.6446 / **0.7774**
WCT [19]/feathers	0.5194 / 0.5816	0.5267 / 0.6324	**0.5822** / **0.6536**	0.6091 / 0.6952	0.5196 / 0.7466	**0.6636** / **0.7893**
WCT [19]/sketch	0.4356 / 0.4742	0.5158 / 0.6203	**0.5930** / **0.6560**	0.5301 / 0.6887	0.4944 / 0.7321	**0.6683** / **0.7839**
WCT [19]/wave	0.5060 / 0.5271	0.4760 / 0.6046	**0.5319** / **0.6268**	0.5819 / 0.6969	0.4843 / 0.7293	**0.6330** / **0.7756**
Fast-neural-style [24]/princess	**0.6468** / 0.6445	0.5170 / 0.6186	0.5965 / **0.6513**	0.6834 / 0.7119	**0.8708** / **0.9212**	0.6344 / 0.7688
Fast-neural-style [24]/udnie	0.5312 / 0.6081	0.4799 / 0.6027	**0.5634** / **0.6422**	0.5951 / 0.7166	0.5495 / 0.7628	**0.6172** / **0.7733**
DBL [11]/expertA	0.2798 / 0.5182	0.4663 / 0.6232	**0.5201** / **0.6475**	0.2857 / 0.6027	0.3843 / 0.6833	**0.4806** / **0.7159**
DBL [11]/expertB	0.1257 / 0.4442	0.3622 / 0.5718	**0.4173** / **0.5968**	0.1499 / 0.5376	0.2792 / 0.6295	**0.4091** / **0.6772**
Intrinsic [2]/reflectance	0.5635 / 0.6353	0.7313 / 0.7384	**0.7717** / **0.7591**	0.7419 / 0.8027	0.7832 / 0.8757	**0.8497** / **0.8832**
Intrinsic [2]/shading	0.5896 / 0.6585	0.7489 / 0.7394	**0.7926** / **0.7512**	0.7059 / 0.8136	0.6623 / 0.8137	**0.8495** / **0.8750**
CycleGAN [51]/photo2ukiyoe	0.2913 / 0.5066	0.4091 / 0.5822	**0.4639** / **0.6056**	0.3492 / 0.6148	0.3659 / 0.6729	**0.5179** / **0.7267**
CycleGAN [51]/photo2vangogh	0.4415 / 0.5742	0.4302 / 0.5922	**0.5271** / **0.6324**	0.4966 / 0.6748	0.3969 / 0.6873	**0.5884** / **0.7569**
Colorization [49]	0.2445 / 0.4772	0.4253 / 0.5958	**0.4806** / **0.6201**	0.3195 / 0.5988	0.3573 / 0.6656	**0.4880** / **0.7132**
Colorization [21]	0.1216 / 0.4304	0.3492 / 0.5601	**0.4079** / **0.5828**	0.1749 / 0.5443	0.2693 / 0.6216	**0.4189** / **0.6802**
Sum	7.0374 / 8.9148	7.9148 / 9.9340	**8.8928** / **10.3542**	8.2096 / 10.8313	7.9506 / 11.7570	**9.7219** / **12.2482**

Table 2: Comparison of the IOU tracking results for the TDMS-Net and the two compared approaches.

Task	DAVIS		
	Bonneel *et al.* [4]	Lai *et al.* [27]	**TDMS-Net**
WCT [19]/antimono	**0.5319**	0.5301	0.5250
WCT [19]/feathers	0.5061	0.5169	**0.5291**
WCT [19]/candy	0.5276	0.5139	**0.5350**
WCT [19]/feathers	0.5011	**0.5118**	0.5053
WCT [19]/sketch	0.4680	0.4929	**0.5073**
WCT [19]/wave	0.4829	**0.4975**	0.4953
Fast-neural-style [24]/princess	**0.5126**	0.4913	0.4991
Fast-neural-style [24]/udnie	**0.5201**	0.4859	0.4990
DBL [11]/expertA	0.5737	0.5848	**0.5864**
DBL [11]/expertB	0.5795	0.5863	**0.5869**
Intrinsic [2]/reflectance	0.5172	0.5321	**0.5459**
Intrinsic [24]/shading	**0.5207**	0.5073	0.4884
CycleGAN [51]/photo2ukiyoe	0.5393	**0.5395**	0.5318
CycleGAN [51]/photo2vangogh	0.4415	0.4302	**0.5271**
Colorization [49]	0.5409	**0.5651**	0.5604
Colorization [21]	0.5533	**0.5732**	0.5531
Sum	8.3947	8.4821	**8.4885**

The Adam [25] optimizer is adopted to train our model with the learning rate is initially set to $1e-4$ with weight decay. Training the TDMS-Net model takes about 60 hours (for 100K iterations) on a workstation with a single NVIDIA 2080Ti GPU card, which is on par with the previous works [4, 27] under the same experimental setting.

For the hyper-parameters in TDMS-Net, λ_{fs} and λ_p are empirically set to 100 and 10, respectively, which are recommended in [27]. Besides, λ_{fd} is set to 0.5 and the λ_w is set to 1, which is following the proposed tuning scheme in [19]. In the latter experiment, we also provide the detailed parameter sensitiveness analysis on the four hyper-parameters.

Evaluation Metrics. Similar to previous works [4, 27], we measure the metric of *temporal consistency* on the output videos, which reflects the temporal smoothness of the generated video. Besides, we also propose another useful metric of *composited spatio-temporal consistency* to explicitly assess the spatial similarity of the frames in the generated video. The definitions of the two metrics are as follows:

Temporal Consistency. The flow warping error is utilized to measure the temporal consistency of the output video sequences:

$$\mathcal{D}_{to} = \sum_{t=2}^{T} \frac{m_{t-1}^{t} \odot \|v_o^{t-1} - F(v_o^t, f_t^{t-1})\|_1}{(T-1)\sum_{i,j} m_{t-1}^{t}(i,j)}, \tag{10}$$

where the f_t^{t-1} is the forward flow which is estimated by the FlowNet. The $m_{t-1}^{t}(i,j)$ represents the pixel of motion mask at the spatial position (i,j). The T is the number of frames in the video. We use the f_t^{t-1} to backward warp the v_o^t and regard the warping error between v_o^{t-1} and the warped v_o^t as the temporal consistency of the frames. Finally, the average warping error over the entire sequence is defined as the temporal consistency of this video.

Moreover, to show how much our model has been improved compared with the original results, we also compute the temporal consistency of the pre-processed video and use the increment ratio to be the final metric:

$$\mathcal{D}_t = \frac{\mathcal{D}_{tp} - \mathcal{D}_{to}}{\mathcal{D}_{tp}}, \tag{11}$$

with $\mathcal{D}_{tp}$ being the temporal consistency of the pre-processed video.

Composited Spatio-temporal Consistency. As mentioned in [27], the improvement of temporal consistency may only indicate that the model has overly smoothed the contents and decrease the perceptual similarity with the pre-processed videos. Therefore, we proposed another metric to evaluate the spatial similarity of the results.

Inspired by the video tracking task which aims to track the target area in the video sequence by giving the label of the first frame, we leverage the available tracking model to evaluate the

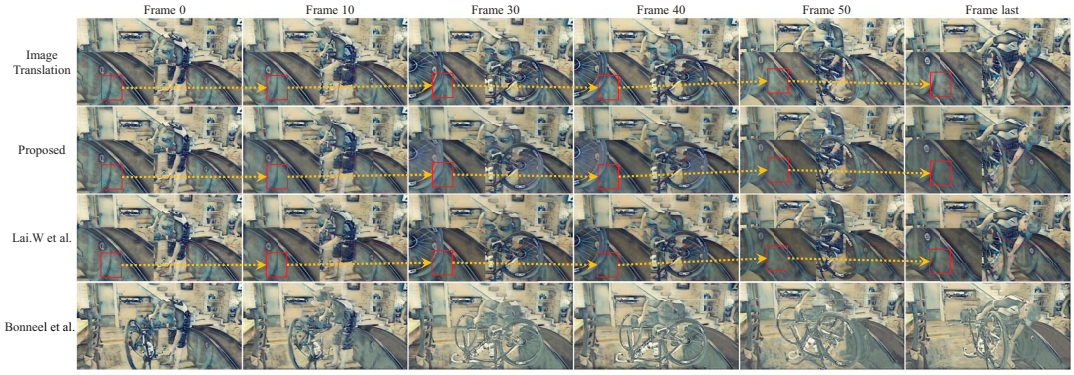


Figure 5: Visualized translated frames by the proposed method and the compared method.

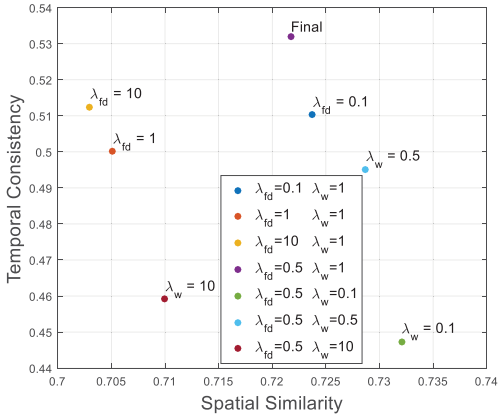


Figure 6: The grid search results of the hyper parameters.

results produced by our model. Simply given the segmentation label of the first frame and the v_o, the tracking model can predict the segmentation mask of the rest frames. It is expected that the higher spatial similarity the output has, the higher tracking IOU it will obtain. Considering the time cost and the accuracy, the SiamMask [36] model is chosen as the tracking network in our experiment.

Therefore, we take the average of the $\mathcal{D}_{IOU}$ (value of IOU) and $\mathcal{D}_{LPIPS}$ (value of LPIPS) to measure the spatial similarity of the output video sequences.

$$\mathcal{D}_s = \frac{1 - \mathcal{D}_{LPIPS} + \mathcal{D}_{IOU}}{2}. \tag{12}$$

Specifically, we use the LPIPS (with the SqueezeNet [20]) to measure the perceptual distance of the processed video $\boldsymbol{v_p}$ and output video $\boldsymbol{v_o}$. Notably, the videos in VIDEVO do not have the segmentation label, so we directly use the $\mathcal{D}_{LPIPS}$ instead of the $\mathcal{D}_s$ to represent the spatial similarity. Finally, the composited spatio-temporal consistency is defined as the average of the temporal consistency $\mathcal{D}_t$ and spatial similarity $\mathcal{D}_s$ ($1 - \mathcal{D}_{LPIPS}$ for VIDEVO).

4.2 Comparison with Existing Methods

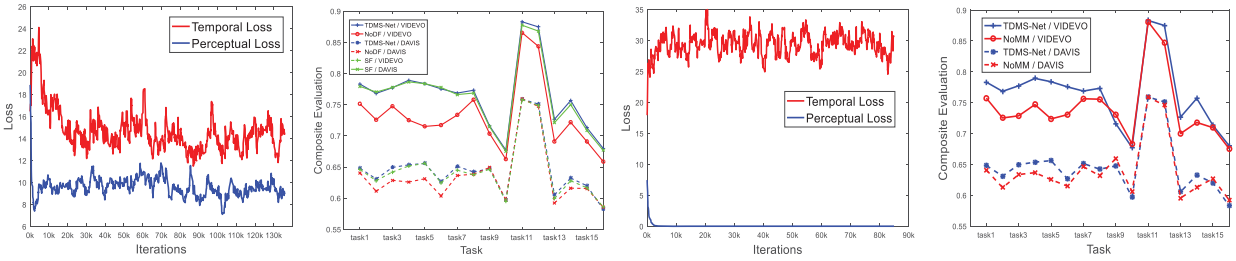
We compare TDMS-Net approach with two state-of-the-art methods Bonneel *et al.* [4] and Lai *et al.* [27] on the DAVIS and VIDEVO datasets. Following the settings in [27], all three methods are evaluated on 8 different tasks in total: (1) Style Transfer [24, 29]; (2) Enhancement [11]; (3) Domain Transfer [51]; (4) Colorization [21]; (5) Intrinsic Decomposition [2]; (6) Automatic white balancing [17]; (7) Image harmonization [3]; and (8) Image dehazing [14]. For all methods, average score of all videos is used as the final score.

The Table 1 shows the quantitative evaluation of both the temporal consistency and composited spatio-temporal consistency of our proposed method and the counterparts under different image-based pre-processing algorithms. We can observe that our proposed method outperforms the two counterparts almost in all tasks for the two evaluation metrics. In particular, for the summation score of all tasks, the TDMS-Net achieves the best result. The overall comparison in Table 1 shows the superiority of our model on handling different conditions in various applications compared to the two counterparts.

Furthermore, we compare the three methods from the visual tracking aspect, where the IOU result is adopted as the measurement. Since the VIDEVO dataset does not have the segmentation label, we report the IOU results of three methods under different applications on the DAVIS dataset. It can be consistently seen that our TDMS-Net obtains better IOU results on majority of the tasks than the other two methods and it achieves the best summation score of all tasks. Thus, it again verifies that TDMS-Net can obtain higher spatial similarity of the processed videos. The visualized comparisons with Bonneel *et al.* [4] and Lai *et al.* [27] are shown in Fig. 5. Notably, when the occlusion occurs in a large region, their method fails due to the lack of a long-term temporal constraint. In contrast, TDMS-Net dramatically reduces the temporal flickering while maintaining the perceptual similarity with the processed videos.

4.3 Further Analysis

Parameter Sensitiveness. TDMS-Net model contains four hyper-parameters, as aforementioned in Section 4.1, the optimal values of hyper-parameters λ_{fs} and λ_p have been tuned and recommended in [27]. In this experiment, to understand the relationship between the dual flow weight λ_{fd} and motion mask weight λ_w, we use one of the styles (*i.e.*, wave) from the WCT transfer method [19] for evaluation. Actually, λ_{fd} and λ_w play a trade-off between the results



(a) The loss tendency of TDMS-Net. (b) Effect of the dual flow loss. (c) The loss tendency of NoRefine-Net. (d) Effect of motion mask.

Figure 7: The subfigures are: the sketch loss tendency of TDMS-Net (a) and TDMS-Net without refining mask learning module (c); comparisons of the model with (b) and without (d) dual flow (motion mask) learning module.

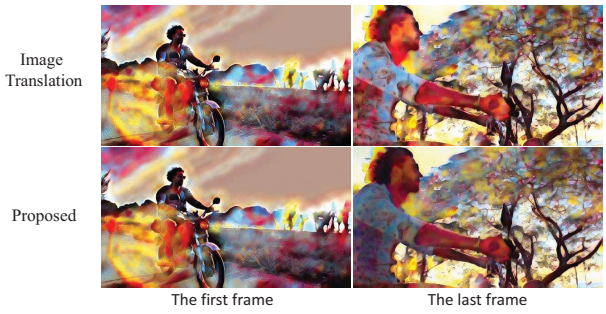


Figure 8: The failure cases of our model, the first row is the pre-processed frames and the second row is our result.

of the temporal consistency and spatial similarity. Therefore, we use the grid search scheme to explore their impact on the final results of the two metrics, which is illustrated in Fig. 6. We can observe that larger values of λ_{fd} will have better temporal consistency and inferior spatial similarity, while λ_w has a saddle point that reaches a good balance between the two metrics. It can be seen that the finally chosen values $\lambda_{fd} = 0.5$ and the $\lambda_w = 1$ provide the best trade-off between the two metrics.

Tendency of Training Loss. During the experiment, we find that the loss value reflects the learning process of our network. As shown in Fig. 7 (a), there's a trade-off between temporal consistency and perceptual similarity. In the early training, the model wants to minimize the content perceptual loss and prefers to output the frames which are identical to the v_p. Once the model learns how to balance these two losses, the model will slightly sacrifice the perceptual similarity for the temporal consistency. Finally our network can balance these two properties and produce ideal results.

Impacts of Different Components in TDMS Module. We also conduct experiments to show the impacts of different modules of our three-stream network. Firstly, as shown in Fig. 7 (b), we remove the dual flow learning module. In those tasks which have more local changes, such as style transfer, the model with dual flow loss can produce more satisfying results. However, in those tasks which enhance the whole frame globally, such as colorization, dual flow contributes less promotion. In addition that, as the video length increases (video in VIDEVO is longer than that in DAVIS), the effect of dual flow increases. The experiment also shows that the soft flow loss $\mathcal{L}_f^{d(1)}$ is more important than $\mathcal{L}_f^{d(2)}$. Secondly, we remove the motion mask learning module, independently. As shown in Fig. 7 (d), the model without motion mask has limited information of the previous frame, therefore, it's harder to produce coherent video sequences. Lastly, we remove the refining mask learning module. We tried many sets of hyperparameters but failed to train it. As shown in Fig. 7 (c), the model without refining mask converges quickly as the model learns to output the identical pre-processed frame v_p instead of learning to reconstruct it.

Finally, we investigate the failure cases of our proposed model. Fig. 8 shows a typical unsatisfying video result produced by our model. As our model basically leverage sufficient information from previous frames during the mask synthesizing process, it may cause a problem especially when the object in the generated frame looks dark in the first few frames and gradually becomes bright.

5 CONCLUSION

In this paper, we proposed the Temporal Denoising Mask Synthesis Network (TDMS-Net) model to solve the flicking problem which caused by using the image-based algorithm to process the frame sequence independently in videos. With the help of dual flow loss, our TDMS-Net is effective to distinguish whether the unstable noise comes from the foreground motion objects or the background area. Meanwhile, the soft flow, motion mask and the refining mask are jointly embedded in TDMS to produce more coherent frames. Using the widely-adopted metric of temporal consistency and our newly proposed composited spatio-temporal consistency, the proposed TDMS-Net model outperforms the two state-of-the-art blind temporal consistency methods on a diverse set of applications and various types of videos.

6 ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (61976049, 61872064 and 61632007); the Fundamental Research Funds for the Central Universities under Project (ZYGX2019Z015) and the Sichuan Science and Technology Program, China (2019ZDZX0008, 2019YFG003 and 2020YFS0057).

REFERENCES

[1] Simon Baker, Daniel Scharstein, J. P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski. 2011. A Database and Evaluation Methodology for Optical Flow. *International Journal of Computer Vision* 92 (2011), 1–31.

[2] Sean Bell, Kavita Bala, and Noah Snavely. 2014. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)* 33 (2014), 1–12.

[3] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. 2015. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision* 51 (2015), 22–45.

[4] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. 2015. Blind Video Temporal Consistency. *ACM Transactions on Graphics (TOG)* 34 (2015).

[5] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. 2004. High accuracy optical flow estimation based on a theory for warping. In *European conference on computer vision*. 25–36.

[6] Andrés Bruhn, Joachim Weickert, and Christoph Schnörr. 2005. Lucas/Kanade Meets Horn/Schunck: Combining Local and Global Optic Flow Methods. *International Journal of Computer Vision* 61 (2005), 211–231.

[7] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. 2012. A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 611–625.

[8] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. 2017. Coherent online video style transfer. (2017), 1105–1114.

[9] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. 2015. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 2758–2766.

[10] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2019. Temporal cycle-consistency learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1801–1810.

[11] Michaël Gharbi, Jiawen Chen, Jonathan T. Barron, Samuel W. Hasinoff, and Frédo Durand. 2017. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)* 36 (2017), 118:1–118:12.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. (2014), 2672–2680.

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

[14] Kaiming He, Jian Sun, and Xiaoou Tang. 2010. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence* 33 (2010), 2341–2353.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[16] Berthold K. P. Horn and Brian G. Schunck. 1981. Determining Optical Flow. *Artif. Intell.* 17 (1981), 185–203.

[17] Eugene Hsu, Tom Mertens, Sylvain Paris, Shai Avidan, and Frédo Durand. 2008. Light mixture estimation for spatially varying white balance. In *ACM SIGGRAPH 2008 papers*. 1–7.

[18] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. 2017. Real-Time Neural Style Transfer for Videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7044–7052.

[19] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.

[20] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size. *CoRR* abs/1602.07360 (2016).

[21] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2016. Let there be color! Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (ToG)* 35 (2016), 1–11.

[22] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2462–2470.

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5967–5976.

[24] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 694–711.

[25] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR 2015*.

[26] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[27] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. 2018. Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 170–185.

[28] Manuel Lang, Oliver Wang, Tunc Aydin, Aljoscha Smolic, and Markus Gross. 2012. Practical temporal consistency for image-based graphics applications. *ACM Transactions on Graphics (ToG)* 31 (2012), 1–8.

[29] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal style transfer via feature transforms. In *Advances in neural information processing systems*. 386–396.

[30] Etienne Mémin and Patrick Pérez. 1998. Dense estimation and object-based segmentation of the optical flow with robust techniques. *IEEE Transactions on Image Processing* 7 (1998), 703–719.

[31] Simon Niklaus and Feng Liu. 2018. Context-Aware Synthesis for Video Frame Interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1701–1710.

[32] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic Image Synthesis With Spatially-Adaptive Normalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2337–2346.

[33] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. 2016. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 724–732.

[34] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. 2015. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1164–1172.

[35] Jingkuan Song, Yuyu Guo, Lianli Gao, Xuelong Li, Alan Hanjalic, and Heng Tao Shen. 2019. From Deterministic to Generative: Multimodal Stochastic RNNs for Video Captioning. *IEEE Transactions on Neural Networks and Learning Systems* 30, 10 (2019), 3047 – 3058.

[36] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. 2019. Fast online object tracking and segmentation: A unifying approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[37] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8798–8807.

[38] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. 2019. Learning Correspondence From the Cycle-Consistency of Time. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2566–2576.

[39] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. 2018. Occlusion aware unsupervised learning of optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4884–4893.

[40] Andreas Wedel, Daniel Cremers, Thomas Pock, and Horst Bischof. 2009. Structure- and motion-adaptive regularization for high accuracy optic flow. In *International Conference on Computer Vision*. 1663–1668.

[41] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. 2013. DeepFlow: Large displacement optical flow with deep matching. In *Proceedings of the IEEE international conference on computer vision*. 1385–1392.

[42] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li. 2019. Ternary Adversarial Networks With Self-Supervision for Zero-Shot Cross-Modal Retrieval. *IEEE Transactions on Cybernetics* (2019), 1–14.

[43] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li. 2017. Learning Discriminative Binary Codes for Large-scale Cross-modal Retrieval. *IEEE Trans. Image Processing* 26, 5 (2017), 2494–2507.

[44] Wenhan Yang, Jiaying Liu, and Jiashi Feng. 2019. Frame-Consistent Recurrent Video Deraining With Dual-Level Flow. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1661–1670.

[45] Chun-Han Yao, Chia-Yang Chang, and Shao-Yi Chien. 2017. Occlusion-aware Video Temporal Consistency. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017,*. 777–785.

[46] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In *The IEEE International Conference on Computer Vision (ICCV)*. 2868–2876.

[47] Jason J. Yu, Adam W. Harley, and Konstantinos G. Derpanis. 2016. Back to Basics: Unsupervised Learning of Optical Flow via Brightness Constancy and Motion Smoothness. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 3–10.

[48] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. 2019. Exploiting temporal consistency for real-time video depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 1725–1734.

[49] Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful image colorization. In *European conference on computer vision*. Springer, 649–666.

[50] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 586–595.

[51] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.