



A Temporal Consistency Enhancement Algorithm Based on Pixel Flicker Correction

Junfeng Meng^(✉), Qiwei Shen^(✉), Yangliu He^(✉), and Jianxin Liao^(✉)

Beijing University of Posts and Telecommunications, Beijing, China
{mengjunfeng, shenqiwei, heyangliu, liaojianxin}@bupt.edu.cn

Abstract. When the image algorithm is directly applied to the video scene and the video is processed frame by frame, an obvious pixel flickering phenomenon is happened, that is the problem of temporal inconsistency. In this paper, a temporal consistency enhancement algorithm based on pixel flicker correction is proposed to enhance video temporal consistency. The algorithm consists of temporal stabilization module TSM-Net, optical flow constraint module and loss calculation module. The innovation of TSM-Net is that the ConvGRU network is embedded layer by layer with dual-channel parallel structure in the decoder, which effectively enhances the information extraction ability of the neural network in the time domain space through feature fusion. This paper also proposes a hybrid loss based on optical flow, which sums the temporal loss and the spatial loss to better balance the dominant role of the two during training. It improves temporal consistency while ensuring better perceptual similarity. Since the algorithm does not require optical flow during testing, it achieves real-time performance. This paper conducts experiments based on public datasets to verify the effectiveness of the pixel flicker correction algorithm.

Keywords: short video creation · deep learning · temporal consistency · optical flow

1 Introduction

Recently, the advances of deep neural networks (DNNs) have led to the rapid development of image processing. Convolutional neural networks [1] (CNNs) take an input image and convert it into the desired output image. This technology has played an important role in image enhancement [2], style transfer [3], image translation [4]. It is nontrivial to extend CNN-based methods to video because of the objective factors such as video datasets and computational constraints. To solve this problem, video is normally processed frame by frame with image algorithms. However, it often leads to undesired inconsistent results in output video with applying image algorithms independently, which is manifested as serious

pixel flickering between frames. Therefore, enhancing temporal consistency aims at eliminating this flickering phenomenon.

The mainstream strategies are divided into task-dependent algorithms and post-processing algorithms. Task-dependent algorithms process each frame independently, relying on stabilization modules or loss functions to maintain temporal consistency. The network architecture of this kind of algorithm has a high degree of coupling and low flexibility. At the same time, the stabilization module, as part of the algorithm, needs to be jointly trained with the task network. While enhancing the temporal consistency, it influences the effect of the original task. Another algorithm is called post-processing algorithm, which is blind to the image algorithm. The principle is to perform secondary processing on the result of the image algorithm. Compared with task-dependent algorithms, it is able to handle multiple tasks. Therefore, post-processing algorithm is the research frontier to solve the problem of temporal consistency. It takes adjacent video frames as input during training and uses optical flow information to achieve consistency.

However, the effect of post-processing algorithms is still insufficient. The reason for this problem is the insufficient ability of temporal stability module to perceive temporal series information. The obtained inter-frame temporal correlation is not rich enough, which leads to the lack of sufficient temporal information when the decoder restores the image.

In this work, a temporal consistency enhancement algorithm is proposed, which can remove flickering in videos. Since the algorithm is blind to the specific image algorithm, it is suitable for various scenarios such as video style transfer, video defogging, and video super-resolution. The algorithm adopts a network architecture based on optical flow constraints, which is divided into temporal stability module TSM-Net, optical flow constraint module and loss calculation module. The optical flow is only calculated during training, while testing is not, which effectively guarantees a high processing speed.

We make the following contributions in this work: (1) We propose a new temporal stabilization module TSM-Net. The module embeds the ConvGRU network layer-by-layer in the decoder to enhance the decoder’s perception of temporal correlations when restoring images. The single-channel serial structure is extended to a dual-channel parallel structure, balancing the guiding role of temporal information and spatial information on the deconvolution layer. At the same time, multi-scale features are fused between the encoder and decoder in a skip connection manner. (2) We propose a new hybrid loss. The loss is divided into long-term and short-term temporal loss and spatial loss. The long-term temporal loss calculates the temporal difference between video frames with interval u . Training model by minimizing the loss function to enhance temporal consistency and ensure perceptual similarity.

2 Related Work

2.1 Task-Dependent Algorithms

Such algorithms normally embed a temporal stabilization module into a deep neural network and retrain the network model with an optical flow-based loss function [5]. Gupta et al. [6] proposes a recurrent neural network for style transfer. The network does not require optical flow during testing and is able to generate time-stable stylized video in real time. Liu et al. [7] adopts a strategy of partial convolution to suppress the jitter phenomenon by using contextual semantic information and correlation. Liu et al. [8] proposes a framework which consists of prediction network, stylized network and loss network. The prediction network and the stylized network extract style information and style information respectively. The loss network is used to train the prediction network and the stylized network. Wei et al. [9] designs a multi-instance normalization block to learn styles, while improving temporal consistency with ConvLSTM [10]. Huang et al. [5] proposes a hybrid loss function that utilizes the content information, style information and temporal information of the input image to improve the stabilization effect of the video. Using the results of salient object segmentation and depth estimation, Liu et al. [11] proposes depth consistency loss and long-term and short-term temporal loss for object occlusion. Xu et al. [12] proposes the frame difference loss function, defined as the spatial distance between the stylized frame and the original frame. The performance of the proposed loss function is excellent. Wang et al. [13] proposes to regularize the temporal series to better balance the spatial and temporal series characteristics, and deal with the complex changes and violent motion in the video.

2.2 Post-processing Algorithms

Post-processing algorithms. Such algorithms use image algorithms to preprocess the video, and then use post-processing techniques to rectify the output. Bonneel et al. [14] proposes a post-processing algorithm that is independent of specific image algorithms. This method takes the original video frame and the video frame processed by the image algorithm as input, and solves the problem of gradient domain optimization by training the model by minimizing the temporal distortion error between frames. On this basis, Yao et al. [15] further improves the effect by estimating optical flow occlusion by selecting key frames. However, the above methods all have high limitations and cannot be applied to the field of image generation, such as style transfer, image defogging, super-resolution. In view of the above shortcomings, Lai et al. [16] regards the video temporal consistency problem as a learnable task and proposes a deep neural network framework. The algorithm takes the raw video and the processing result of the image algorithm as input, and generates an output video with stable temporal. Zhou et al. [17] regards the video temporal consistency problem as a denoising problem and proposes a temporal denoising mask synthesis network (TDMS-Net). It jointly predicts motion masks, optical flow, and refinement masks to synthesize stable frames.

In addition, Lei et al. [18] proposes a general framework that draws on the idea of Deep Video Prior (DVP) to mine the temporal information contained in the original videos. Although the algorithm can solve the flickering problem of image denoising and super-resolution, it needs to train each test video separately during testing, which cannot achieve the effect of practical application.

3 Methodology

Figure 1 shows the temporal consistency enhancement algorithm with temporal stabilization module (TSM-Net), loss calculation module and optical flow constraint module. The input is the Image Package, which consists of four images I_{t-1} , O_{t-1} , I_t and P_t , corresponding to the current time t , representing the input frame and output frame at time $t - 1$, and the input frame and image algorithm processing result at time t .

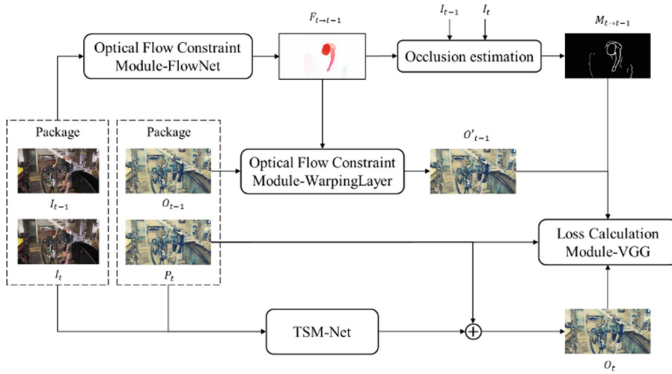


Fig. 1. The Framework of the Temporal Consistency Enhancement Algorithm

During training, the algorithm enhances the ability of temporal consistency enhancement with optical flow constraint module and ensures the perceptual similarity with loss calculation module. Since optical flow is not needed during testing, the calculation amount of the process is greatly reduced. As a result, it has a faster processing speed in practical applications and can achieve real-time performance. This paper employs warp error and learned perceptual metric to evaluate the quality of the output video. Experiments show that the proposed algorithm achieves a good balance between enhancing temporal consistency and perceptual similarity.

3.1 Temporal Stability Module

Temporal Stabilization Module is the core of the algorithm to extract features and restore images. As shown in Fig. 2, it consists of network layers such as Con-layer, Upsample-Layer, ConvGRU-Layer, and residual modules. Among them,

\oplus represents the element-wise addition operation, \odot represents the concat operation, and \rightarrow represents the transfer direction of image features. The architecture is divided into a contraction path and an expansion path. The left shrinking path is used as an encoder to obtain contextual information. It extracts features layer by layer through convolution operations. The symmetrical expansion path on the right is used as a decoder to restore the image. As the rich temporal information between the original video frames, the adjacent video frames I_{t-1} and I_t are input to the module. The features are extracted by downsampling and fused with other features, which can help the decoder to output images related to the timing of the previous moment. The purpose of inputting O_{t-1} and P_t into this module is to extract rich spatial detail features. Summing the output features with P_t can fuse the semantic information of shallow and deep layers, making the algorithm more robust.

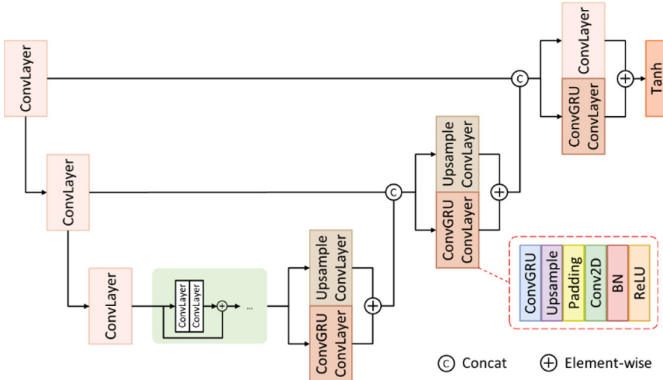


Fig. 2. The Structure of TSM-Net

In order to avoid transmitting the low-level features to decoder and bringing noise to the color space, the image is divided into two branches before the downsampling. As shown in Fig. 3, the two branches here do not set up two layers of convolution for Input1 and Input2 separately, but use the same convolution layer.

The decoder uses three deconvolution layers to restore the image, and the kernel size is opposite to the corresponding encoder, which are 3×3 , 3×3 and 7×7 respectively and the stride is 1. After extracted features by the convolutional neural network, the size of the output normally becomes smaller. Deconvolution layer can restore the image to its original size, enabling small to large resolution mapping.

ConvGRU-Layer consists of ConvGRU and Upsample-Layer, that are connected in series to capture the spatiotemporal correlation of video. The effectiveness of this idea has been confirmed in the Matting [19] task. Different from this design, this paper expands the single-channel serial network into a dual-channel parallel network, and the input of each layer is multi-scale fusion features.

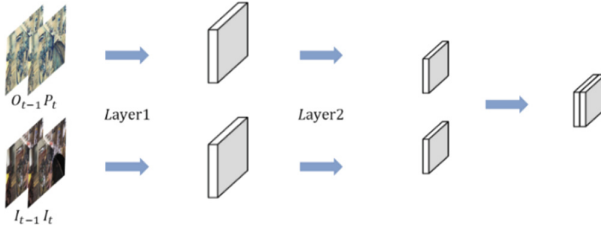


Fig. 3. The Shared Parameters

Features are input into ConvGRU-Layer and deconvolutional layers for upsampling, respectively, and then summed by pixel-wise.

3.2 Hybrid Loss Function

The hybrid loss includes spatial loss and temporal loss. The spatial loss is used to supervise the network restoration and the detailed features of the video frames processed by the image algorithm, and the temporal loss is used to reduce the temporal difference between the video frames.

Spatial Loss. We compute the content similarity between O_t and P_t . The spatial features of O_t and P_t are extracted by the pretrained VGG-19 classification network, and then the content loss is calculated, which is defined as:

$$\mathcal{L}_p = \sum_{t=2}^T \sum_{i=1}^N \|\varphi_i(O_t) - \varphi_i(P_t)\|_1 \quad (1)$$

where $\varphi_i(\cdot)$ represents the feature vector extracted by the i th layer VGG-19 network, and N represents the number of VGG network layers, T represents the total number of video sequence samples. Here we calculate the loss of the entire video frame with L1 loss function. Since the first frame does not participate in the calculation, the spatial loss calculation time is 2 to T .

Short-Term Temporal Loss. We compute L1 loss between O_t and O'_{t-1} to measure the warping error between the output frames:

$$\mathcal{L}_t^s = \sum_{t=2}^T M_{t \rightarrow t-1} \odot \|O_t - O'_{t-1}\|_1 \quad (2)$$

where O'_{t-1} represents the frame warped by optical flow, $M_{t \rightarrow t-1}$ represents the optical flow confidence of each pixel, \odot represents element-wise operation process. During the experiment, we select FlowNet2.0 to calculate the optical flow. The optical flow only needs to be used once during training, while is not required during testing.

Long-Term Temporal Loss. Although short-term loss can ensure the temporal stability between adjacent frames, it is slightly insufficient for long sequences

of video frames. An intuitive way to enhance long-term stability is to compute the temporal loss between O_t with all other frames in the sequence, but it is computationally expensive. Another method is to calculate the error between O_t with the first frame as a long-term loss. Although this method can reduce the amount of calculation, it causes a larger error because the interval is too long. To this end, a new long-term loss is proposed, it is defined as:

$$\mathcal{L}_t^l = \sum_{t=u}^T M_{t \rightarrow t-u} \odot \|O_t - O'_{t-u}\|_1 \quad (3)$$

where u represents the interval, T represents the total number of video sequence samples. During training, $s = 5$, $T = 25$. The advantage is that the calculation amount and the correction effect are well balanced.

Hybrid Loss. The hybrid loss combines the spatial loss and the temporal loss, which is defined as:

$$\mathcal{L} = \lambda_p \mathcal{L}_p + \lambda_t^s \mathcal{L}_t^s + \lambda_t^l \mathcal{L}_t^l \quad (4)$$

where λ_p , λ_t^s , λ_t^l represent the weights corresponding to the spatial loss \mathcal{L}_p , the short-term temporal loss \mathcal{L}_t^s and the long-term temporal loss \mathcal{L}_t^l .

4 Experiments

4.1 Datasets

The dataset used for training is divided into the DAVIS-2017 dataset [20] and the open dataset collected by Lai et al. DAVIS-2017 is a VOS dataset in the field of instance segmentation with more video frame sequences multiple moving objects and motion types. It includes 60 video clips for training and 30 video clips for validation. There are 10731 frames of 150 videos in total, including 4209 frames of 60 videos in the training set and 2023 frames of 30 videos in the validation set. However, the videos in the DAVIS dataset are usually short in length, on average less than 3s. To this end, Lai et al. collected 100 additional high-quality videos from Video.net [21], of which 80 videos are used for training and 20 videos are used for testing. We select the style transfer algorithm WCT [22], the image enhancement algorithm [2] and the image colorization algorithm of Zhang et al. [23] and Bell et al. [24] to preprocess the dataset.

4.2 Model Training and Inference

In terms of dataset, the image size is set to 192×192 ; the length of the input video clip is 25; the weights of the loss are 100, 100 and 10 respectively; the video frame interval u is set to 5; the scale used to extract image features is set to 4; the initial learning rate is 0.0001, which decays every 20 cycles with a decay rate of 0.5; the minimum learning rate is 0.00001; the optimizer is Adam; the batchsize is set to 2. One Tesla P100 GPU is used in training.

4.3 Evaluation Metrics

This paper is to correct the flickering of pixels between frames and generate a stable output video, while ensuring the perceptual similarity between each output frame. To this end, the following two metrics are used to measure the temporal stability and perceptual similarity of videos.

Temporal Stability. The temporal stability of the video is measured based on the optical flow distortion error between two frames:

$$\varepsilon_{wrap}(O_t, O'_{t+1}) = \frac{1}{\sum_{i=1}^{W \times H \times C} M_t^{(i)}} \sum_{i=1}^{W \times H \times C} M_t^{(i)} \odot \|O_t^{(i)} - O'_{t+1}^{(i)}\|_1 \quad (5)$$

where O'_{t+1} represents the distorted frame obtained by FlowNet at time $t + 1$, O_t represents the output frame at time t , M_t is an occlusion estimation, \odot represents pixel product.

$$E_{wrap}(O) = \frac{1}{T-1} \sum_{t=1}^{T-1} \varepsilon_{wrap}(O_t, O_{t+1}) \quad (6)$$

Equation 6 expresses the mean value of all distortion errors from the first frame to the T-1 frame as the temporal error of entire video sequence.

Perceptual Similarity. The pretrained VGG network can extract image features and be used as an effective loss computing network to supervise TSM-Net to generate more realistic images. This performance has been verified in multiple vision tasks. On this basis, Zhang et al. [23] et al. proposes a perceptual metric and introduced a new human perceptual similarity judgment dataset. This indicator can effectively correspond to human perceptual judgment, and has good performance in tasks such as super-resolution frame interpolation and image deblurring. In this paper, this indicator is used to measure the perceptual distance between the processed video P and the output video O. The equation is expressed as:

$$D_{perceptual}(P, O) = \frac{1}{T-1} \sum_{t=2}^{T-1} \delta_{wrap}(O_t, P_t) \quad (7)$$

where $\delta(\cdot)$ represents the calibration model LPIPS, similar to the algorithms proposed by Lai, the algorithm proposed in this paper also excludes the first frame when calculating the perceptual distance.

4.4 Experimental Results and Analysis

The temporal series correction network TSM-Net proposed in this paper is a post-processing technology, which has nothing to do with the specific algorithm used to preprocess the video frame sequence. In the training process, a variety of

task algorithms are selected to generate the training set, including style transfer, image dehazing, super-resolution, etc. In the model effect evaluation, only the wave-style model in the style transfer algorithm WCT is selected to generate the pre-processed video based on DAVIS-2017, and then the performance of the model is tested as a test set. The total data set used for evaluation is 30 video clips, 2023 frames. The final output video is obtained by inputting the above video frames into TSM-Net.

100 models are saved during training. We calculate the corresponding E_{warp} and $D_{perceptual}$, and draw a scatter plot (as shown in Fig. 4). It shows that the two indicators are in opposition. As the E_{warp} decreases, the $D_{perceptual}$ increases. Videos with high perceptual similarity have obvious flickering problem, so the balance of the two play an important role in the actual effect of the model. According to Lai, it is found that: when $r = \lambda_p \setminus \lambda_t$ is greater than 0.1, the corrected output video still has obvious pixel flickering; while r is less than 0.1, the influence of supervision of temporal loss is greater and the output video becomes very blurry at this point. Therefore, if r is fixed at 0.1, it can balance the two phenomena better, and this setting is also suitable for other tasks.

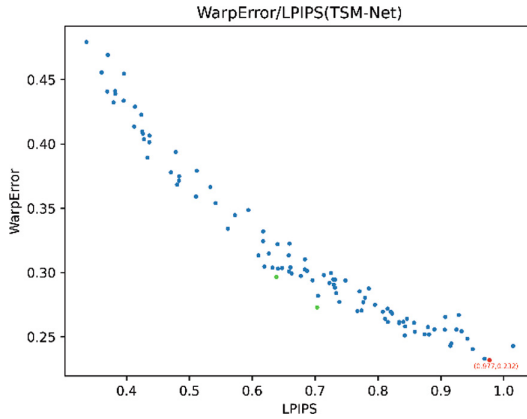


Fig. 4. Test Result of TSM-Net

The red dot in Fig. 4 corresponds to the model with the highest temporal consistency, and E_{warp} is 0.232 and $D_{perceptual}$ is 0.977. The two green dots correspond to models that combine temporal consistency and perceptual similarity. These two models well balance the phenomenon of pixel flickering and video blur, and are suitable for application. The red dots in Fig. 4 also correspond to the models with the highest temporal consistency. Its E_{warp} is 0.302 and $D_{perceptual}$ is 0.745. Comparing the results of TSM-Net and the algorithm proposed by Lai, the temporal consistency is improved by 23.18%. The two green dots in Fig. 4 are compared with Fig. 5. The perceptual similarity of the model with better comprehensive effect is in the range of 0.6 ~ 0.7, while the E_{warp} of TSM-Net

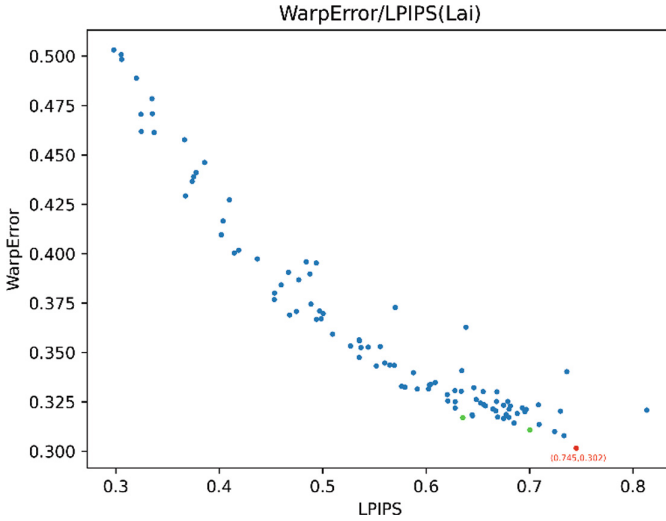


Fig. 5. Test Result of the Algorithm Proposed by Lai et al.

is significantly lower than that of Lai, indicating that the perceived similarity unchanged and the correction effect of TSM-Net is better than that of Lai.



Fig. 6. Comparison of Visualization

As shown in Fig. 6, each row has five pictures taken from the output frame sequence in chronological order. From top to bottom, the original input video, the preprocessed video, the output video of Lai’s algorithm and the output video of TSM-Net are displayed in order. We can clearly see that the background of the preprocessed video has obvious pixel flickering. Although there has been a certain improvement in temporal consistency of the algorithm of Lai et al., the correction effect of TSM-Net is even better. As shown in Table 1, the results of

temporal consistency metrics are presented. The first two columns represent the results of the preprocessed videos as controls. Compared with the preprocessed video, the test results of the algorithm proposed by Lai have been improved to a certain extent, but the effect is limited. The last two columns show the test results of TSM-Net. Our algorithm performs the best for different image processing tasks whether it is the DAVIS-2017 dataset or the videvo dataset (Table 2).

Table 1. Comparison of State of Art Algorithm

Task	Vp (10-2)		Lai (10-2)		Ours (10-2)	
	DAVIS	Videvo	DAVIS	Videvo	DAVIS	Videvo
WCT/wave	0.660	0.577	0.302	0.211	0.232	0.143
WCT/feathers	0.780	0.698	0.335	0.250	0.243	0.171
WCT/sketch	0.612	0.540	0.240	0.188	0.176	0.127
WCT/candy	0.865	0.703	0.380	0.235	0.309	0.177
WCT/asheville	1.039	0.886	0.478	0.308	0.353	0.204
WCT/antimono	0.591	0.476	0.257	0.150	0.215	0.124
FastNeural/princess	1.162	1.100	0.472	0.429	0.363	0.282
FastNeural/udnie	0.429	0.325	0.187	0.110	0.157	0.085
DBL/expertA	0.222	0.176	0.117	0.096	0.108	0.076
DBL/expertB	0.162	0.122	0.103	0.072	0.097	0.057
Intrinsic/reflectance	0.357	1.183	0.083	0.075	0.077	0.058
Intrinsic/shading	0.392	0.793	0.081	0.052	0.074	0.044
CycleGAN/ukiyo	0.314	0.271	0.178	0.132	0.152	0.099
CycleGAN/vangogh	0.438	0.397	0.228	0.185	0.188	0.123
Average	0.573	0.590	0.246	0.178	0.196	0.126

Table 2. Experimental effects of technologies

Technologies	E_warp	Decrease
Shared weight	0.302	2.71% (parameter)
Hybrid loss	0.263	15.34%(E_warp)
Dual-channel parallel	0.276	11.21%(E_warp)
TSM-Net	0.232	25.30%(E_warp)

4.5 Ablation Experiments

This section mainly introduces the effectiveness of the related technologies used in TSM-Net. As shown in Table 1, the E_{warp} of the algorithm proposed by Lai

is 0.302, which is used as the control experiment result. In this paper, the ConvGRU network is introduced into the temporal stabilization module, which is connected in series between the residual module and the decoder to enhance its ability to obtain temporal information. When the E_{warp} remains unchanged, the parameter quantity is reduced by 9.9%, which verifies the effectiveness of ConvGRU. The network model of ConvGRU is used as a control experiment.

The shared weight strategy adopted by the first two layers in the downsampling effectively reduces the complexity of the network, and the E_{warp} of the test results can remain unchanged. It means that the shared weight does not affect the temporal stability effect, and can also complete the feature extraction goal. At the same time, it avoids the passing of shallow information to the decoder.

This paper proposes a new hybrid loss including spatial loss and temporal loss. The spatial loss integrates the multi-scale features extracted by the VGG network, which is divided into short-term temporal loss and long-term temporal loss. The long-term temporal loss with a calculation time interval of u enhances the stability of the network for long video clips, and the E_{warp} of the test results is reduced by 15.34%, which verifies the effectiveness of the hybrid loss.

The ConvGRU network is connected in each Upsampling-Layer, and the single-channel network is expanded into a dual-channel parallel network. Then the output of the deconvolution layer and the output of the ConvGRU network layer are fused. The E_{warp} of the test results is reduced by 11.21%, which verifies the effectiveness for temporal stability enhancement.

Applying all the above techniques to TSM-Net, the E_{warp} is 0.2318, which is 25.30% lower than the algorithm proposed by Lai. It proves the positive effect of various techniques on enhancing temporal consistency.

5 Conclusion

In this paper, we propose a temporal consistency enhancement algorithm, which can solve the problem of pixel flicker. Since the algorithm is blind to the specific image algorithm, it is suitable for style transfer, defogging, and super-resolution while ensuring high processing speed. The network adopts a network architecture based on optical flow constraints, which is divided into temporal stabilization module TSM-Net, loss calculation module and optical flow constraint module. The algorithm only calculates the optical flow during training, and does not need optical flow during testing, effectively ensuring that the algorithm can be implemented for real-time requirements. Compared with the baseline model, the temporal stabilization module and hybrid loss are innovative. On the one hand, TSM-Net uses the ConvGRU layer-by-layer in the decoder with dual-channel parallel structure. On the other hand, the long-term temporal loss calculates the temporal difference between video frames with interval u . After experimental verification, TSM-Net has greatly improved the ability to enhance temporal consistency, and can balance the perceptual similarity well. Comparing with state-of-the-art methods, our algorithm achieves the best results.

References

1. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998)
2. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. *ACM Trans. Graph. (TOG)* **36**(4), 1–12 (2017)
3. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510 (2017)
4. Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., Yang, M.-H.: Diverse image-to-image translation via disentangled representations. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *ECCV 2018. LNCS*, vol. 11205, pp. 36–52. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_3
5. Huang, H., et al.: Real-time neural style transfer for videos. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 783–791 (2017)
6. Gupta, A., Johnson, J., Alahi, A., Fei-Fei, L.: Characterizing and improving stability in neural style transfer. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4067–4076 (2017)
7. Liu, S., Wu, H., Luo, S., Sun, Z.: Stable video style transfer based on partial convolution with depth-aware supervision. In: *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2445–2453, October 2020
8. Liu, X., Ji, Z., Huang, P., Ren, T.: Real-time arbitrary video style transfer. In: *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*, pp. 1–7, March 2021
9. Gao, W., Li, Y., Yin, Y., Yang, M.H.: Fast video multi-style transfer. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3222–3230 (2020)
10. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: *Advances in Neural Information Processing Systems*, vol. 28 (2015)
11. Liu, Y., Jiang, A., Pan, J., Liu, J., Ye, J.: Deliberation on object-aware video style transfer network with long-short temporal and depth-consistent constraints. *Neural Comput. Appl.* **33**(14), 8845–8856 (2021)
12. Xu, J., Xiong, Z., Hu, X.: Frame Difference-Based Temporal Loss for Video Stylization. arXiv preprint [arXiv:2102.05822](https://arxiv.org/abs/2102.05822) (2021)
13. Wang, W., Yang, S., Xu, J., Liu, J.: Consistent video style transfer via relaxation and regularization. *IEEE Trans. Image Process.* **29**, 9125–9139 (2020)
14. Bonneel, N., Tompkin, J., Sunkavalli, K., Sun, D., Paris, S., Pfister, H.: Blind video temporal consistency. *ACM Trans. Graph. (TOG)* **34**(6), 1–9 (2015)
15. Yao, C.H., Chang, C.Y., Chien, S.Y.: Occlusion-aware video temporal consistency. In: *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 777–785, October 2017
16. Lai, W.-S., Huang, J.-B., Wang, O., Shechtman, E., Yumer, E., Yang, M.-H.: Learning blind video temporal consistency. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *ECCV 2018. LNCS*, vol. 11219, pp. 179–195. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01267-0_11

17. Zhou, Y., Xu, X., Shen, F., Gao, L., Lu, H., Shen, H.T.: Temporal denoising mask synthesis network for learning blind video temporal consistency. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 475–483, October 2020
18. Lei, C., Xing, Y., Chen, Q.: Blind video temporal consistency via deep video prior. *Adv. Neural. Inf. Process. Syst.* **33**, 1083–1093 (2020)
19. Lin, S., Yang, L., Saleemi, I., Sengupta, S.: Robust high-resolution video matting with temporal guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 238–247 (2022)
20. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 724–732 (2016)
21. Beachfront: Stock footage of Videvo (2015). <https://www.videvo.net/>
22. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
23. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 649–666. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_40
24. Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. *ACM Trans. Graph. (TOG)* **33**(4), 1–12 (2014)