Machine Intelligence Research
www.mi-research.net

21(1), February 2024, 46-54
DOI: 10.1007/s11633-023-1447-3

Deep Video Harmonization by Improving Spatial-temporal Consistency

Xiuwen Chen Li Fang Long Ye Qin Zhang

Key Laboratory of Media Audio and Video Ministry of Education, Communication University of China, Beijing 100024, China

Abstract: Video harmonization is an important step in video editing to achieve visual consistency by adjusting foreground appearances in both spatial and temporal dimensions. Previous methods always only harmonize on a single scale or ignore the inaccuracy of flow estimation, which leads to limited harmonization performance. In this work, we propose a novel architecture for video harmonization by making full use of spatiotemporal features and yield temporally consistent harmonized results. We introduce multiscale harmonization by using nonlocal similarity on each scale to make the foreground more consistent with the background. We also propose a foreground temporal aggregator to dynamically aggregate neighboring frames at the feature level to alleviate the effect of inaccurate estimated flow and ensure temporal consistency. The experimental results demonstrate the superiority of our method over other state-of-the-art methods in both quantitative and visual comparisons.

Keywords: Harmonization, temporal consistency, video editing, video composition, nonlocal similarity.

Citation: X. Chen, L. Fang, L. Ye, Q. Zhang. Deep video harmonization by improving spatial-temporal consistency. *Machine Intelligence Research*, vol.21, no.1, pp.46–54, 2024. http://doi.org/10.1007/s11633-023-1447-3

1 Introduction

Video composition is one of the most common operations in video editing tasks. However, generating a composite video by combining the foreground of one video and the background of another video may look unrealistic due to the incompatible appearances from the different shooting environments, photo equipment, etc. To address this issue, video harmonization can be used to ensure the realism of the generated video. In general, video harmonization aims to adapt the appearances of the foreground video to make it compatible with the new background.

For video harmonization tasks, there are two main challenges: One is to attain realistic harmonized results by leveraging background information, and the other is to exploit temporal information between consecutive harmonized frames. To obtain realistic harmonized results, image harmonization, which aims to obtain harmonized images by adjusting foreground images, has been extensively explored. Conventional methods[1–7] address the harmonization problem by transferring statistics of handcrafted features between foreground and background regions, such as color and texture. However, these methods mostly work in some simple cases, causing unreliable results when the appearances are vastly different. In recent years, many deep learning-based approaches have been proposed for generating harmonized images. Existing methods usually use UNet-like structures as backbones, cooperating with various strategies, such as attention mechanisms[8–11], semantic information[12–14], illumination exchange[15–17], to achieve decent results. Among them, Hao et al.[9] and Hang et al.[10] have proven the efficiency of nonlocal similarity information in harmonization tasks, by exploiting self-similarity across pictures in the bottleneck of their networks. Unlike their methods, we consider leveraging multiscale features to reconstruct the foreground region, which can effectively capture the correlation among nonlocal patches to make the foreground more consistent with its background.

Temporal consistency is a frequently studied topic in video-related tasks[18–20], and it is the bridge to expand image-to-image methods to process video. Because of massive redundant information in video sequences, there is a high correlation between multiple neighboring frames. Directly applying image harmonization methods to video sequences frame by frame is an easy solution to video harmonization; however, this operation may produce unsightly results that suffer from flickering. Therefore, it is necessary to exploit temporal correlations to achieve temporally consistent video harmonization results. Previous approaches leveraged regional temporal loss[18] or color mapping consistency[19] to generate consistent harmonized frames. Huang et al.[18] used temporal loss, which can

Research Article
Special Issue on Artificial Intelligence for Art
Manuscript received on November 30, 2022; accepted on April 11, 2023
Recommended by Associate Editor Chun-Hua Shen
Colored figures are available in the online version at https://link.springer.com/journal/11633

achieve temporal consistency to a certain extent, but their method has poor performance at harmonization and strict requirements for training datasets. Lu et al.[19] improved their methods, using optical flow to align adjacent frames. However, obtaining accurate optical flow is a challenging task because of inevitable errors in the estimation process, and inaccurate alignment affects the temporal performance. Additionally, Lu et al.[19] proposed a new framework using the lookup table (LUT) of neighboring color mapping to achieve consistency, but their methods operate at the pixel level, which would be unstable.

In this paper, we propose a novel framework that can alleviate flickering artifacts and achieve realistic harmonized results. We find that the key factor is to fuse the corresponding patches in the temporal dimension to fine-tune the current frame. Specifically, first, we encoder composite frames to high-dimension features. Second, we introduce multiscale harmonization by taking advantage of nonlocal self-similarity in multiscale features to obtain harmonized frames and features. Then, for each harmonized framewise feature, we utilize several carefully designed modules, including flow-guided feature propagation modules and foreground temporal aggregators, to model interframe correspondence and effectively aggregate foreground features between adjacent frames in bidirectional ways. Additionally, the feature propagation module aims to utilize both backward and forward propagation schemes to progressively align features of adjacent frames, and a foreground temporal aggregator is used to adaptively aggregate useful information between frames and eliminate alignment errors of warping by estimated optical flow. Finally, we use a decoder to output temporally consistent harmonized results. Generally, our main contributions are summarized as follows:

1) We propose a novel framework for video harmonization, which introduces a multiscale harmonization module and a foreground temporal aggregator with a flow-guided bidirectional propagation strategy, to further improve the performance of both image harmonization and temporal consistency.

2) Compared with other state-of-the-art methods, our method achieves superior performance on the benchmark dataset HYouTube. Extensive ablation studies and visualizations validate the effectiveness of the proposed approach.

2 Related work

2.1 Image harmonization

Conventional methods for image harmonization mainly focus on matching low-level pixel statistics between the foreground and the background, such as color distribution mapping[2, 3, 5, 7], gradient-domain compositing[1, 4], and multiscale statistics matching[6]. Recently, various approaches using deep neural networks have been proposed to improve performance. Tsai et al.[14] and Sofiiuk et al.[13] both introduced semantic information to image harmonization networks and found the efficacy of high-level semantic features. Cun and Pun[8] and Hao et al.[9] proposed effective attention mechanisms for image harmonization. Cong[21, 22] formulated the image harmonization task as foreground-background domain translation. Ling et al.[23] and Zhu et al.[11] explicitly used background style to guide the foreground harmonization, and they designed the RAIN module and self-consistent style contrastive learning scheme to harmonize the images. Guo et al.[16, 24] disentangled composite images into reflectance and illumination for further separate harmonization. Cong et al.[25] and Hu et al.[17], Bao et al.[15] used rendered images and illumination exchange to achieve more authentic results. Cong et al.[26], Liang et al.[27] and Xue et al.[28] focused on high-resolution image harmonization and achieved better harmonization performance with higher efficiency.

2.2 Video harmonization

Deep learning boosts image harmonization with excellent performance and high efficiency, which inspires its transition to video modality. Previous image harmonization methods can be extended to videos if we directly treat every video frame as an independent image. However, owing to the motion and occlusion of foreground objects, directly applying those approaches to an input video may result in temporally inconsistent videos of low visual quality. To suppress flicker results, Huang et al.[18] first trained a convolutional neural network using a pixelwise disharmony discriminator to achieve more realistic harmonized results and then introduced a temporal loss to preserve temporal consistency between consecutive harmonized frames. Lu et al.[19] designed a harmonization network applying the color mapping strategy at the pixel level to achieve temporal consistency. Different from these methods, we introduce a new network for video harmonization that can aggregate multiframe information of foreground regions for better video harmonized results.

2.3 Temporal alignment

Motion compensation is an essential component for most video tasks to handle displacement among frames. Many approaches[20, 29–33] have been developed to enforce temporal consistency in video editing. Ruder et al.[20] employed a temporal loss guided by optical flow for video style transfer. Wang et al.[32] synthesized videos with temporal consistency by training a network to estimate optical flow and applied it to previously generated frames. Ye et al.[33] explicitly leverage temporal information by build-

ing a causal-anticausal, coarse-to-fine iterative scheme. Lang et al.[31] proposed an efficient framework to enforce temporal smoothness across frames. They approximate a global optimization and show very good results for applications such as disparity estimation, depth upsampling or colorization. Bonneel et al.[29] applied curvature-flow smoothing in the space of color transformations to transfer color palettes between videos and demonstrated that it successfully produced temporally consistent results without degrading the video content. Gupta et al.[30] presented a recurrent convolutional network by using previous stylized frame and the current frame as input to produce the stylized current frame. Unlike these methods, we obtain foreground-feature fusion by utilizing neighbor frame information in both forward and backward directions to solve the temporal consistency problem in video harmonization tasks.

3 Our method

3.1 Framework overview

Given a composite video sequences $\{I_i \in \mathbf{R}^{(H\times W\times 3)}, i=1,\cdots,T\}$ as input, with corresponding framewise binary foreground masks $\{M_i \in \mathbf{R}^{(H\times W\times 1)}, i=1,\cdots,T\}$, we aim to generate realistic composite frames that are consistent in both spatial and temporal dimensions. The pipeline of our framework is shown in Fig. 1. First, a feature pyramid network is used as our feature encoder to extract high-dimensional features from the i-th composite frame. Second, the extracted features are handled in a multiscale harmonization module to obtain harmonized features F_i^h. Third, a flow estimation module $E_{b,f}$ is introduced to estimate the bidirectional optical flow of composite frames to model foreground movements. Fourth, guided by optical flow, features F_i^h would be aligned in bidirectional propagation ways. Fifth, to alleviate the misalignment of warping, a foreground temporal aggregator is proposed to fuse the temporal features in foreground regions. Finally, a decoder, which contains several residual blocks and convolutional layers, is used to output temporal harmonization frames $\tilde{I}_i^r$.

3.2 Multiscale harmonization module

Nonlocal self-similarity is one of the important characteristics of natural images and has been proven to be an effective prior for image restoration tasks[34]. For image harmonization, Hao et al.[9] and Hang et al.[10] introduced self-attention mechanisms to calculate nonlocal information and proved the effectiveness of nonlocal information across image regions. However, these methods simply operate on single-scale feature maps. Instead, we consider correlations at multiple scales and exploit nonlocal self-similarity in multiscale features to ensure consistency between foreground and background regions. As shown in Fig. 2(a), each scale level consists of an upsampling layer, a nonlocal block, and three 3×3 convolution layers in our method. We adopt the nonlocal module proposed in [34], as illustrated in Fig. 2(b), m is the number of channels of the input feature, and l is the number of channels of the intermediate feature. The output at each location in the foreground region is computed by using its $q\times q$ neighborhood.

3.3 Flow-guided bidirectional propagation strategy

Camera or object motion leads to displacement among frames. Directly operating on the nonaligned features may cause substandard performance. This is because convolutions, as local operations, have relatively small receptive fields and are inefficient in aggregating the information from corresponding locations. Thus, it is important to adopt operations that have a sufficiently large receptive field to aggregate information from distant spatial locations. To better establish interframe correspond-

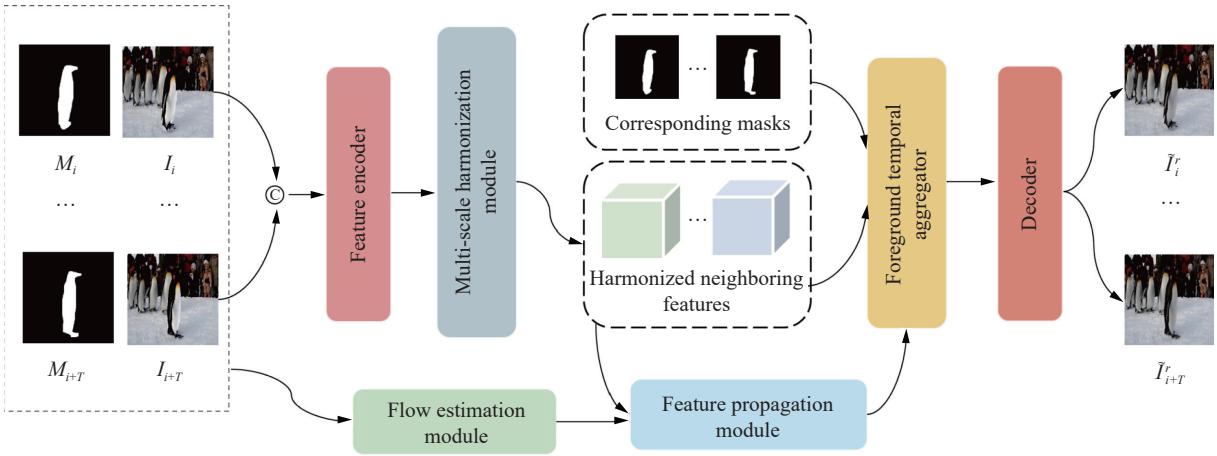

Fig. 1 Overview of the proposed framework. It consists of 1) a feature encoder, 2) a multiscale harmonization module, 3) a flow estimation module, 4) a feature propagation module, 5) a foreground temporal aggregator, and 6) a decoder.

ence, we adopt simple flow-based alignment methods before aggregation. By computing optical flow and warping on feature maps, we extract motion features to guide our model to focus on moving foreground objects. In addition, to make full use of multiframe information, we utilize a bidirectional propagation strategy, in which features can independently propagate in both forward and backward directions of time. In general, our methods use composite frames I_i to estimate bidirectional optical flow and perform warping on the features F_i^h in both forward and backward directions in the temporal dimension. Formally, we have

$$\begin{aligned} S_i^{\{b,f\}} &= S(I_i, I_{i\pm t}) \\ \bar{F}_i^{h\{b,f\}} &= W(F_{i\pm t}^{h\{b,f\}}, S_i^{\{b,f\}}) \end{aligned} \tag{1}$$

where $\bar{F}_i^{h\{b,f\}}$ denotes aligned features, S and W denote the flow estimation and spatial warping modules, respectively. In addition, b denotes backward flow, and f denotes forward flow.

3.4 Foreground temporal aggregator

Optical flow can generate high-quality output but is computationally expensive, and its estimation may suffer from inevitable errors, which may affect the reconstruction performance. Therefore, dynamically aggregating neighboring frames at the feature level is indispensable for effective and efficient aggregation. To address the above problem, we propose an attention module to attain more accurately aligned features by assigning feature-level aggregation weights to the neighboring foreground, as shown in Fig. 3. We first use 1×1 convolutions to map feature maps of both the current frame and its adjacent frames to embedding spaces and then calculate the similarity in embedding spaces. Theoretically, compared with the misaligned pixels, the correctly aligned pixels in adjacent frames are more similar to the current frame, which should be given more attention. For each frame $i \in [1:T]$, the similarity distance can be obtained by

$$f(\bar{F}_{i-t}^h, \bar{F}_i^h) = \text{sigmoid}(\theta(\bar{F}_{i-t}^h)^{\mathrm{T}}, \varphi(\bar{F}_i^h)) \tag{2}$$

where θ and φ denote two embeddings.

Then, we apply the attention map A and mask M_i, and we obtain the attention feature map as $\hat{F}_i^h$, which is then concatenated with the feature map of the current frame. Finally, we feed concatenated features into the decoder and obtain output frames. We have

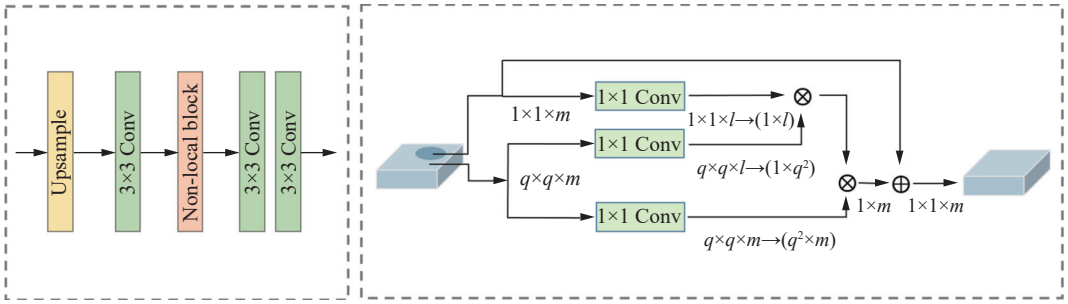

Fig. 2 Architecture of the proposed operation for harmonization, including: (a) detailed operations of each scale level harmonization; (b) nonlocal block we used

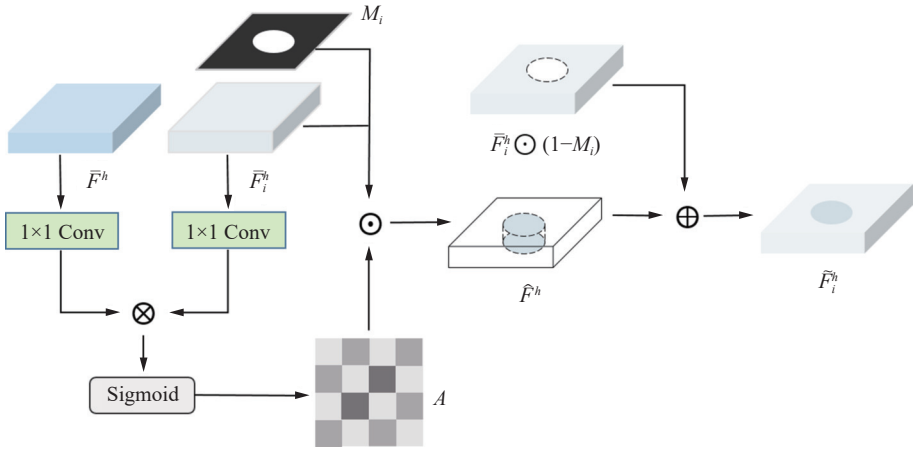

Fig. 3 Architecture of the proposed foreground temporal aggregator

$$\tilde{F}_i^h = M_i \odot \bar{F}_i^h \odot f(\bar{F}_{i-t}^h, \bar{F}_i^h) + (1 - M_i) \odot \bar{F}_i^h \quad (3)$$

where $\odot$ denotes the elementwise multiplication.

4 Experiment

4.1 Dataset statistics

To evaluate the proposed model and make fair comparisons with state-of-the-art approaches, we use the HYouTube dataset proposed by Lu et al.[19] The HYouTube dataset contains 3 194 training and 636 test video sequences with foreground masks. The dataset is synthesized from the large-scale video dataset YouTube-VOS2018 and is constructed by using abundant 3D color lookup table to adjust the appearance of foregrounds to make frames incompatible with backgrounds. The video length in HYouTube is 20 frames. Furthermore, to eliminate the gap between real composite videos and synthetic composite videos, HYouTube contains another 100 real composite videos via copy-and-paste for testing. Each real composite video has 20 frames.

4.2 Implementation details

The network takes 5 (or 3) consecutive frames as input. Similar to CO_2Net[19], we use iS^{2}AM[13] as our basic module. We use two nonlocal blocks in two scale levels, and q is set to 4 as the default. We use pretrained SPyNet[35] as our flow estimation module for more accurate flow estimation. In the decoder, we implement 5 residual blocks with channel size 32.

We apply a two-stage training strategy. The first stage is to obtain harmonized images frame by frame and train the network by using the foreground-normalized (FN)-MSE loss proposed by [13], and the second stage concentrates on obtaining temporal consistency results, with L1 loss for training. We train our network on a single RTX 3 090 GPU for 150 epochs and adopt the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The initial learning rates are set to 5×10^{-4} and reduced by a factor of 10 at epochs 85 and 125. The batch size is 16 (or 24), and the size of composite frames is 256×256. The input frames are scaled to [0, 1] and normalized with RGB mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225).

4.3 Comparison with existing methods

We compare our method with both image harmonization methods and video harmonization methods. For image harmonization methods, we consider the effectiveness of the proposed algorithm and compare it with existing methods iS^{2}AM[13], RainNet[23], DoveNet[21], and intrinsic image harmonization (IIH)[16], which harmonize videos frame by frame. For the video harmonization methods, we compare with the two existing methods [18] and CO_2Net[19].

For fair comparison with [18] and CO_2Net, we utilize the temporal loss (TL) metric to measure temporal consistency between frames. We also provide fMSE, MSE, PSNR and fSSIM metrics following [19] on the test sets of the HYouTube dataset. Specifically, PSNR and SSIM are frequently used metrics for distortion-oriented image and video assessment. fMSE or fSSIM means only calculating the MSE or SSIM in the foreground region. The results are listed in Table 1. Among the image harmonization methods, our method achieves the best results, and can exactly improve the harmonization performance. As we can see, our method can also improve the temporal consistency for video harmonization, and outperforms the state-of-the-art video harmonization approaches. All these results verify the effectiveness of our method. We also provide some visual comparisons in Fig. 4. It can be seen that our method obtains consistent harmonization results across frames and our harmonized results are closer to ground-truth frames.

Table 1 Comparisons between our method and existing methods on the HYouTube dataset

Methods	fMSE ↓	MSE ↓	PSNR ↑	fSSIM ↑	TL ↓
Composite	1 029.50	151.20	30.14	0.719 7	2.531 5
DoveNet	347.73	47.84	35.06	0.839 2	18.153 3
IIH	333.65	45.91	34.99	0.832 4	3.327 7
RainNet	310.47	42.52	35.49	0.841 1	4.502 4
iS^{2}AM	203.78	28.90	37.38	0.881 7	6.476 5
Huang et al.[18]	198.86	27.81	37.47	0.882 4	6.489 3
CO_2Net	186.71	26.50	37.61	0.882 7	5.112 6
Ours	174.81	24.20	37.89	0.882 6	5.032 8

4.4 Ablation studies

In this section, we conduct ablation studies to evaluate the effectiveness of each component in our methods. The results are shown in Table 2, Rows 1 and 2 present the image harmonization results. Row 1 denotes the results of the harmonization module in iS^{2}AM as our base harmonization module. Row 2 denotes the results of our harmonization method within base harmonization module and nonlocal modules. As we can see, the modification of the base method can generally improve harmonization performance. Row 3 shows the validity of the unidirectional propagation scheme. Compared to Row 6, it can be seen that using bidirectional optical flow can better improve the performance than unidirectional only. Rows 4 and 6 reveal the effectiveness of the pretrained optical flow. By observing Rows 5 and 6, we can find the effectiveness of our foreground temporal aggregator, and

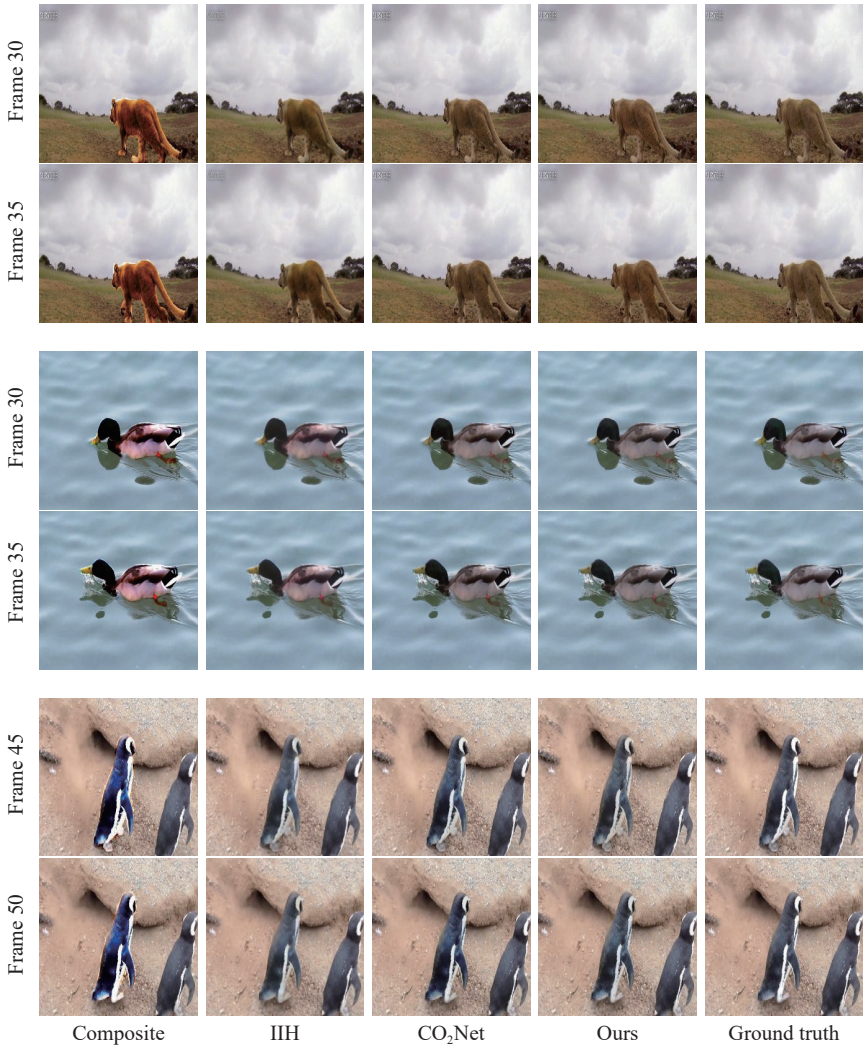

Fig. 4 Visual comparison on HYouTube between our network and other state-of-the-art methods. Zoom in for better visualization.

Table 2 Ablation studies of our framework

	Base	Nonlocal module	Unidirectional propagation	Bidirectional propagation	Pretrained flow	Foreground temporal aggregator	fMSE ↓	TL ↓
1	+						203.78	6.476 5
2	+	+					176.39	6.510 5
3	+	+	+		+	+	179.18	5.615 4
4	+	+		+		+	180.83	5.609 3
5	+	+		+	+		178.44	5.523 0
6	+	+		+	+	+	174.81	5.032 8

our module shows the improvement of temporal consistency.

We conduct the experiment on the number of scale levels of our harmonization module. Our multiscale harmonization module contains three scale levels for placing nonlocal blocks, so we evaluate the effectiveness of this block by using different numbers. We can see from Table 3 that placing a nonlocal block in two scale levels achieves the best results.

In Table 4, we evaluate the performance of using different numbers of input frames. In consideration of computing resources, we conduct experiments on 3, 5 and 7 frames and set the batch size to 10. It can be seen that the larger numbers of frames would get better results.

5 Conclusions

In this paper, we have proposed a new video harmon-

Table 3 Evaluating the effectiveness of multiscale harmonization module with different numbers of scale levels

Numbers	fMSE ↓	MSE ↓	PSNR ↑
0	191.24	25.11	37.40
1	179.89	24.06	37.60
2	176.39	23.32	37.74
3	180.25	24.84	37.62

Table 4 Evaluating the performance of different numbers of input frames

NFrames	fMSE ↓	fSSIM ↑	Time (s) ↓
3	184.03	0.880 5	0.168
5	179.14	0.881 6	0.187
7	176.27	0.882 3	0.298

ization network, combining multiscale harmonization, flow-guided feature propagation, and foreground temporal aggregation to address the temporally consistent harmonization of composite videos. By exploiting multiscale nonlocal self-similarity and foreground temporal aggregation, our method can achieve appearance consistency and temporal consistency of composite video sequences. Experimental results have shown that our method outperforms state-of-the-art methods in both quantitative and visual performance on the HYouTube dataset.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 62001432), and the Fundamental Research Funds for the Central Universities, China (Nos. CUC18LG024 and CUC22JG001).

Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

References

[1] P. Pérez, M. Gangnet, A. Blake. Poisson image editing. *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 313–318, 2003. DOI: 10.1145/882262.882269.

[2] E. Reinhard, M. Adhikhmin, B. Gooch, P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, vol. 21, no. 5, pp. 34–41, 2001. DOI: 10.1109/38.946629.

[3] E. Reinhard, A. Oguz Akyuz, M. Colbert and C. E. Hughes, M. O'Connor. Real-time color blending of rendered and captured video. In *Proceedings of Interservice/Industry Training, Simulation, and Education Conference*, Article number 1502, 2004.

[4] Y. Jia, J. Sun, C. K. Tang, H. Y. Shum. Drag-and-drop pasting. *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 631–637, 2006. DOI: 10.1145/1141911.1141934.

[5] J. F. Lalonde, A. A. Efros. Using color compatibility for assessing image realism. In *Proceedings of the 11th IEEE International Conference on Computer Vision*, IEEE, Rio de Janeiro, Brazil, 2007. DOI: 10.1109/ICCV.2007.4409107.

[6] K. Sunkavalli, M. K. Johnson, W. Matusik, H. Pfister. Multi-scale image harmonization. *ACM Transactions on Graphics*, vol. 29, no. 4, Article number 125, 2010. DOI: 10.1145/1778765.1778862.

[7] S. Xue, A. Agarwala, J. Dorsey, H. E. Rushmeier. Understanding and improving the realism of image composites. *ACM Transactions on Graphics*, vol. 31, no. 4, Article number 84, 2012. DOI: 10.1145/2185520.2185580.

[8] X. D. Cun, C. M. Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing*, vol. 29, pp. 4759–4771, 2020. DOI: 10.1109/TIP.2020.2975979.

[9] G. Q. Hao, S. Iizuka, K. Fukui. Image harmonization with attention-based deep feature modulation. In *Proceedings of the 31st British Machine Vision Conference*, 2020.

[10] Y. C. Hang, B. Xia, W. M. Yang, Q. M. Liao. SCS-Co: Self-consistent style contrastive learning for image harmonization. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 19678–19687, 2022. DOI: 10.1109/CVPR52688.2022.01909.

[11] Z. Y. Zhu, Z. Zhang, Z. Lin, R. Q. Wu, Z. Chai, C. L. Guo. Image harmonization by matching regional references, [Online], Available: https://arxiv.org/abs/2204.04715, 2022.

[12] X. Q. Ren, Y. F. Liu. Semantic-guided multi-mask image harmonization. In *Proceedings of the 17th European Conference on Computer Vision*, Springer, Tel Aviv, Israel, pp. 564–579, 2022. DOI: 10.1007/978-3-031-19836-6_32.

[13] K. Sofiiuk, P. Popenova, A. Konushin. Foreground-aware semantic representations for image harmonization. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, Waikoloa, USA, pp. 1619–1628, 2021. DOI: 10.1109/WACV48630.2021.00166.

[14] Y. H. Tsai, X. H. Shen, Z. Lin, K. Sunkavalli, X. Lu, M. H. Yang. Deep image harmonization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 2799–2807, 2017. DOI: 10.1109/CVPR.2017.299.

[15] Z. Y. Bao, C. J. Long, G. Fu, D. Q. Liu, Y. Z. Li, J. M. Wu, C. X. Xiao. Deep image-based illumination harmonization. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 18521–18530, 2022. DOI: 10.1109/CVPR52688.2022.01799.

[16] Z. H. Guo, H. Y. Zheng, Y. F. Jiang, Z. R. Gu, B. Zheng. Intrinsic image harmonization. *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 16362–16371, 2021. DOI: 10.1109/CVPR46437.2021.01610.

[17] Z. Y. Hu, N. E. Nsampi, X. Wang, Q. Wang. NeurSF: Neural shading field for image harmonization, [Online], Available: https://arxiv.org/abs/2112.01314, 2021.

[18] H. Z. Huang, S. Z. Xu, J. X. Cai, W. Liu and S. M. Hu. Temporally coherent video harmonization using adversarial networks. *IEEE Transactions on Image Processing*, vol. 29, pp. 214–224, 2020. DOI: 10.1109/TIP.2019.2925550.

[19] X. Y. Lu, S. C. Huang, L. Niu, W. Y. Cong, L. Q. Zhang. Deep video harmonization with color mapping consistency. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, Vienna, Austria, pp. 1232–1238, 2022.

[20] M. Ruder, A. Dosovitskiy, T. Brox. Artistic style transfer for videos. In *Proceedings of the 38th DAGM German Conference on Pattern Recognition*, Springer, Hannover, Germany, pp. 26–36, 2016. DOI: 10.1007/978-3-319-45886-1_3.

[21] W. Y. Cong, J. F. Zhang, L. Niu, L. Liu, Z. X. Ling, W. Y. Li, L. Q. Zhang. DoveNet: Deep image harmonization via domain verification. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 8391–8400, 2020. DOI: 10.1109/CVPR42600.2020.00842.

[22] W. Y. Cong, L. Niu, J. F. Zhang, J. Liang, L. Q. Zhang. BargainNet: Background-guided domain translation for image harmonization. In *Proceedings of IEEE International Conference on Multimedia and Expo*, Shenzhen, China, pp. 1–6, 2021. DOI: 10.1109/ICME51207.2021.9428394.

[23] J. Ling, H. Xue, L. Song, R. Xie, X. Gu. Region-aware adaptive instance normalization for image harmonization. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 9357–9366, 2021. DOI: 10.1109/CVPR46437.2021.00924.

[24] Z. H. Guo, D. S. Guo, H. Y. Zheng, Z. R. Gu, B. Zheng, J. Y. Dong. Image harmonization with transformer. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 14850–14859, 2021. DOI: 10.1109/ICCV48922.2021.01460.

[25] W. Y. Cong, J. Y. Cao, L. Niu, J. F. Zhang, L. Q. Zhang. Deep image harmonization by bridging the reality gap, [Online], Available: https://arxiv.org/abs/2102.17104, 2022.

[26] W. Y. Cong, X. H. Tao, L. Niu, J. Liang, X. S. Gao, Q. H. Sun, L. Q. Zhang. High-resolution image harmonization via collaborative dual transformations. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 18449–18458, 2022. DOI: 10.1109/CVPR52688.2022.01792.

[27] J. T. Liang, X. D. Cun, C. M. Pun. Spatial-separated curve rendering network for efficient and high-resolution image harmonization. In *Proceedings of the 17th European Conference on Computer Vision*, Springer, Tel Aviv, Israel, pp. 334–349, 2022. DOI: 10.1007/978-3-031-20071-7_20.

[28] B. Xue, S. H. Ran, Q. Chen, R. F. Jia, B. Q. Zhao, X. Tang. DCCF: Deep comprehensible color filter learning framework for high-resolution image harmonization. In *Proceedings of the 17th European Conference on Computer Vision*, Springer, Tel Aviv, Israel, pp. 300–316, 2022. DOI: 10.1007/978-3-031-20071-7_18.

[29] N. Bonneel, K. Sunkavalli, S. Paris, H. Pfister. Example-based video color grading. *ACM Transactions on Graphics*, vol. 32, no. 4, Article number 39, 2013. DOI: 10.1145/2461912.2461939.

[30] A. Gupta, J. Johnson, A. Alahi, F. F. Li. Characterizing and improving stability in neural style transfer. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 4087–4096, 2017. DOI: 10.1109/ICCV.2017.438.

[31] M. Lang, O. Wang, T. O. Aydin, A. Smolic, M. Gross. Practical temporal consistency for image-based graphics applications. *ACM Transactions on Graphics*, vol. 31, no. 4, Article number 34, 2012. DOI: 10.1145/2185520.2185530.

[32] T. C. Wang, M. Y. Liu, J. Y. Zhu, G. L. Liu, A. Tao, J. Kautz, B. Catanzaro. Video-to-video synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 1152–1164, 2018. DOI: 10.5555/3326943.3327049.

[33] G. Z. Ye, E. Garces, Y. B. Liu, Q. H. Dai, D. Gutierrez. Intrinsic video and applications. *ACM Transactions on Graphics*, vol. 33, no. 4, Article number 80, 2014. DOI: 10.1145/2601097.2601135.

[34] D. Liu, B. H. Wen, Y. C. Fan, C. C. Loy, T. S. Huang. Non-local recurrent network for image restoration. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 1680–1689, 2018. DOI: 10.5555/3326943.3327097.

[35] A. Ranjan, M. J. Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 2720–2729, 2017. DOI: 10.1109/CVPR.2017.291.

Xiuwen Chen received the B. Eng. degree in digital media technology from Communication University of China, China in 2019. She is currently a master student in communication and information systems at Key Laboratory of Media Audio and Video (Communication University of China), Ministry of Education, China.

Her research interests include computer vision and virtual reality.

E-mail: cxw@cuc.edu.cn

ORCID iD: 0000-0002-3507-5446

Li Fang received the B. Eng. degree in electronic information engineering and the M. Sc. degree in communication and information systems from Wuhan University, China in 2010 and 2012, respectively, the Ph. D. degree in communication and information systems from Communication University of China, China in 2017. He is currently an associate professor with Key Laboratory of Media Audio and Video (Communication University of China), Ministry of Education, China. Prior to that, he was a visiting scholar with Ryerson University, Canada.

His research interests include computer vision and virtual reality.

E-mail: lifang8902@cuc.edu.cn (Corresponding author)
ORCID iD: 0000-0001-9963-6110

Long Ye received the B. Eng. degree in electronic engineering from Shandong University, China in 2003, and the M. Sc. and Ph. D. degrees in communication and information systems from Communication University of China, China in 2006 and 2012, respectively. He is currently a professor with State Key Laboratory of Media Convergence and Communication, Communication University of China, China. Prior to that, he was a visiting scholar with Ryerson University, Canada.

His research interests include computer vision, image compression and virtual reality.

E-mail: yelong@cuc.edu.cn
ORCID iD: 0000-0002-3562-5612

Qin Zhang received the Ph. D. degree in engineering from The University of British Columbia, Canada in 1990. He worked as a research and development scientist with EE Department, The University of British Columbia, Canada from 1990 to 1995. In 2004, he served as the Dean of the TCL Industrial Research Institute, China. He is currently a professor with State Key Laboratory of Media Convergence and Communication, Communication University of China, China.

His research interests include computer vision, audio and video processing and virtual reality.

E-mail: zhangqin@cuc.edu.cn

Machine Intelligence Research

Citation: X. Chen, L. Fang, L. Ye, Q. Zhang. Deep video harmonization by improving spatial–temporal consistency. *Machine Inteligence Research*, vol.21, no.1, pp.46–54, 2024. https://doi.org/10.1007/s11633–023–1447–3