Exploring Motion Ambiguity and Alignment for High-Quality Video Frame Interpolation Supplementary Materials

Kun Zhou[1,2] Wenbo Li[3] Xiaoguang Han[1] Jiangbo Lu[2*]
[1]SSE, CUHK-Shenzhen, [2]SmartMore Corporation [3]CUHK
kunzhou@link.cuhk.edu.cn, wenboli@cse.cuhk.edu.hk
hanxiaoguang@cuhk.edu.cn,jiangbo.lu@gmail.com

A. Other Configurations of Our Framework

Common settings. As illustrated in Fig.2, there are four modules in our framework: feature extraction, cross-scale pyramid alignment, attention-based feature fusion, and reconstruction. The number of parameters are 4.28M, 12.52M, 0.29M and 11.80M, respectively, in a total of 28.89M. Following [10], we adopt the residual block [1] as the basic component (shorted as "RB"), which is detailed in Table A.1. In our network, the channel number of convolutions is set to 128. We use $\Rightarrow$ to point out the output of a layer in Tabs. A.1 to A.3.

Input	x
Layer1	Conv(128,128,3,1) + ReLU
Layer2	Conv(128,128,3,1) $\Rightarrow y$
Output	x+y
Params.	0.3M

Table A.1. The structure of the residual block ("RB").

Feature extraction. The structure of the feature extraction module is shown in Table A.2. For a given input frame $I_i \in R^{C\times H\times W}$ $(i = \{-1, 1\})$, we first utilize a convolution to change its channel dimension to 128. Then the feature maps are passed through five residual blocks, resulting in the 0-th level feature F_i^0 of the pyramid representation. Finally, we use two convolutional layers with strides of 2 to generate the downsampled features F_i^1 and F_i^2, respectively.

Input	I_i
Layer1	Conv(3,128,3,1) + ReLU
Layer2	5× RB(128) $\Rightarrow F_i^0$
Layer3	Conv(3,128,3,2) + ReLU $\Rightarrow F_i^1$
Layer4	Conv(3,128,3,2) + ReLU $\Rightarrow F_i^2$
Params.	4.28M

Table A.2. The structure of our feature extraction module.

Reconstruction Table A.3 shows the details of the reconstruction module. The fused intermediate feature F_0 is firstly passed to a sequence of residual blocks for refinement. At last, we use a single convolution without activation to generate the final result I_0.

*Corresponding author

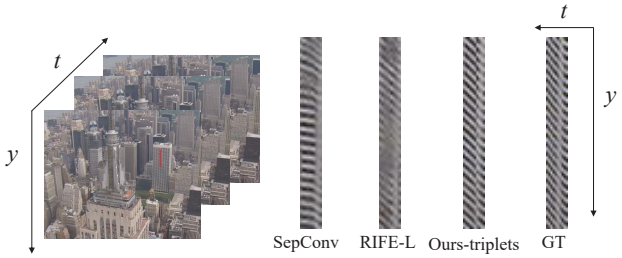


Figure A.1. Temporal consistency analysis of three single-frame VFI approaches. The SepConv [7] and RIFE-L [2] cannot generate continuous signals, while our algorithm shows a smoother transition. The sample comes from Vid4 [5].

Input	F_0
Layer1	40×RB(128)
Layer2	Conv(128,3,3,1) $\Rightarrow \hat{I}_0$
Params.	11.80M

Table A.3. The structure of the reconstruction module.

B. More Results

Temporal consistency. Apart from the quantitative evaluation of PSNR and SSIM, temporal consistency [4, 11] is also an important measure within the realm of video frame interpolation. We compare our method with two representative methods including SepConv [7] and RIFE-L [2] in Fig. A.1. It is observed that SepConv and RIFE-L generate blurry and inconsistent patterns along the time axis, while our method successfully restores the correct and consistent patterns compared with the ground truth.

More visual comparison of TCL. In our paper, we conducted extensive experiments to evaluate the effectiveness of our texture consistency loss (Fig.1, Table 2, Table 3). Specifically, we demonstrate how TCL is useful to train the existing video frame interpolation/extrapolation mod-

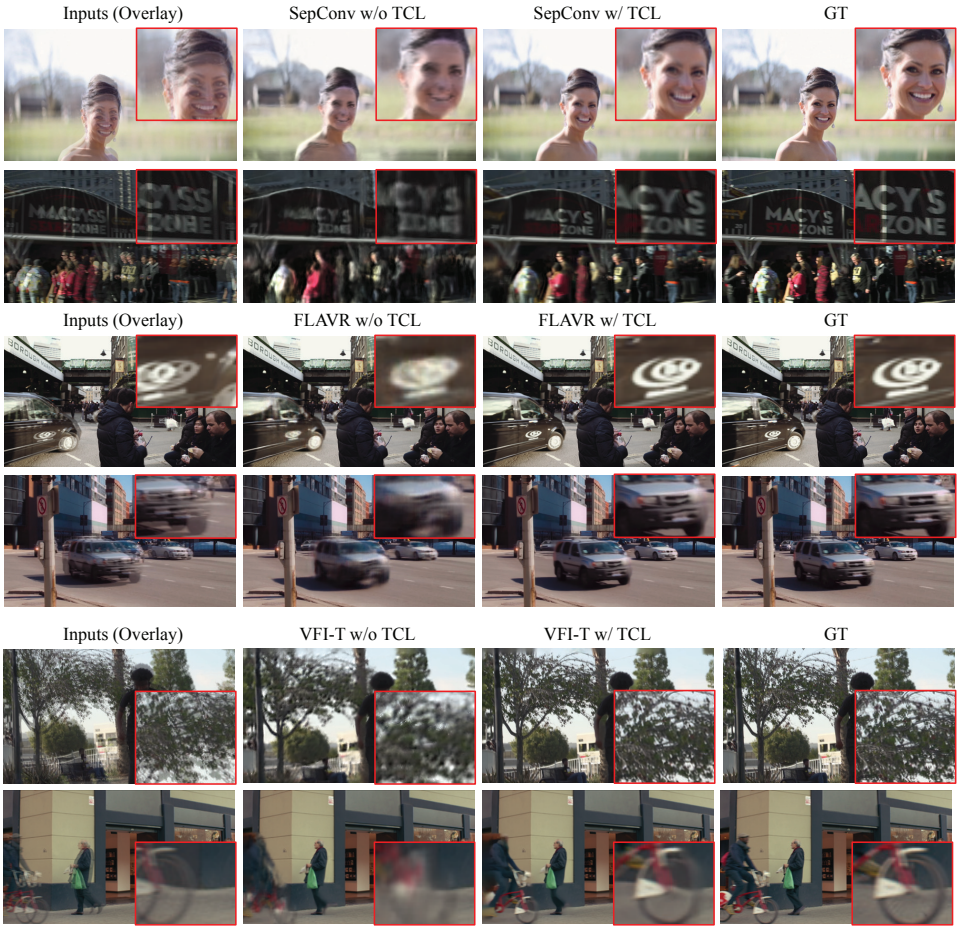


Figure B.2. Visualized results of FLAVR [3] and VFI-T [8] without/with our TCL on Vimeo-Triplets-Test for video frame extrapolation.

els. For instance, SepConv trained with our TCL achieves promising video frame interpolation/extrapolation results on various benchmarks. Additionally, TCL is able to improve the performance of more recent SOTA methods, *i.e.*, FLAVR [3] and VFI-T [8]. As shown in Fig. B.3a, Fig. B.3b and Fig. B.4, we give more visual examples of FLAVR [3] and VFI-T [8], SepConv [7] and our method with/without TCL. It is clear that the proposed TCL is beneficial in hallucinating more plausible structures.

High-resolution and large-motion cases. We test our VFI model on Xiph-4K [1] and show the quantitative comparison in the table below. Our model consistently achieves the **best** performance. Though XVFI [9] is designed for large-motion and high-resolution scenarios (using 286K high-resolution training samples), our model trained on 51K low-resolution data still outperforms it by **3.21dB**. It manifests that our model is capable of interpolating large-motion and high-resolution frames.

Flow-based VFI models + TCL. In this manuscript, we demonstrate how TCL can help existing VFI models. Here,

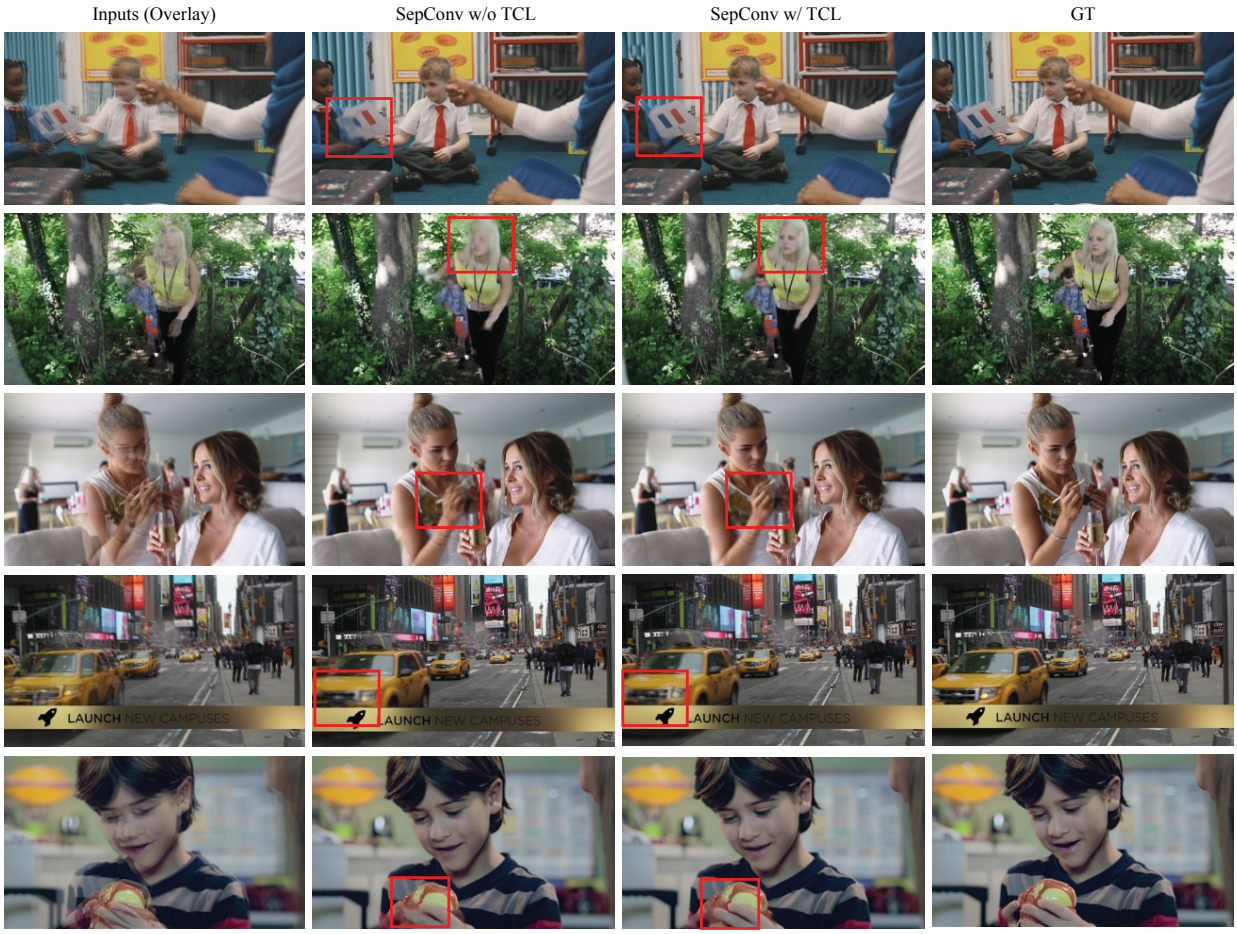
Methods	SepConv	AdaCoF	XVFI	ABME	CURE	Ours
PSNR (dB)	22.83	28.57	28.46	30.74	30.94	**31.67**

we additionally train RIFE [2] with our TCL loss from scratch. The model achieves a **0.19dB** improvement over the officially released model. This experiment illustrates how the proposed texture consistency loss can improve the performance of the flow-based VFI model.

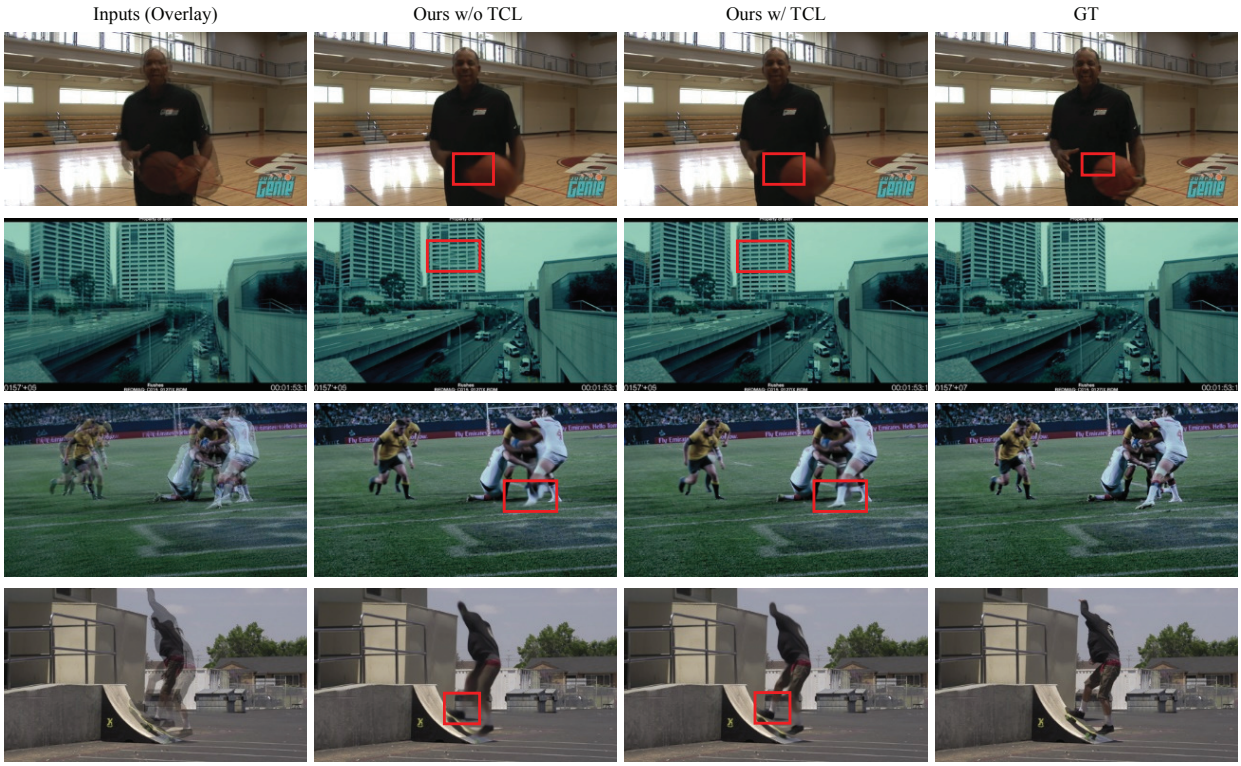
More qualitative results. Here, we provide more visual examples to adequately validate the performance of our approach in Fig. B.4(a-b). It can be observed that the SOTA approaches produce blurry or noisy image contents. In contrast, thanks to the texture consistency loss and guided cross-scale aggregation designs, our proposed approach is capable of interpolating/extrapolating high-quality frames with clearer details and fewer artifacts for the challenging cases (large motion, small objects).

Video result. We also provide a video sequence (named as "comparison.mp4") for visual comparison with VFIformer [6] and FLAVR [3].

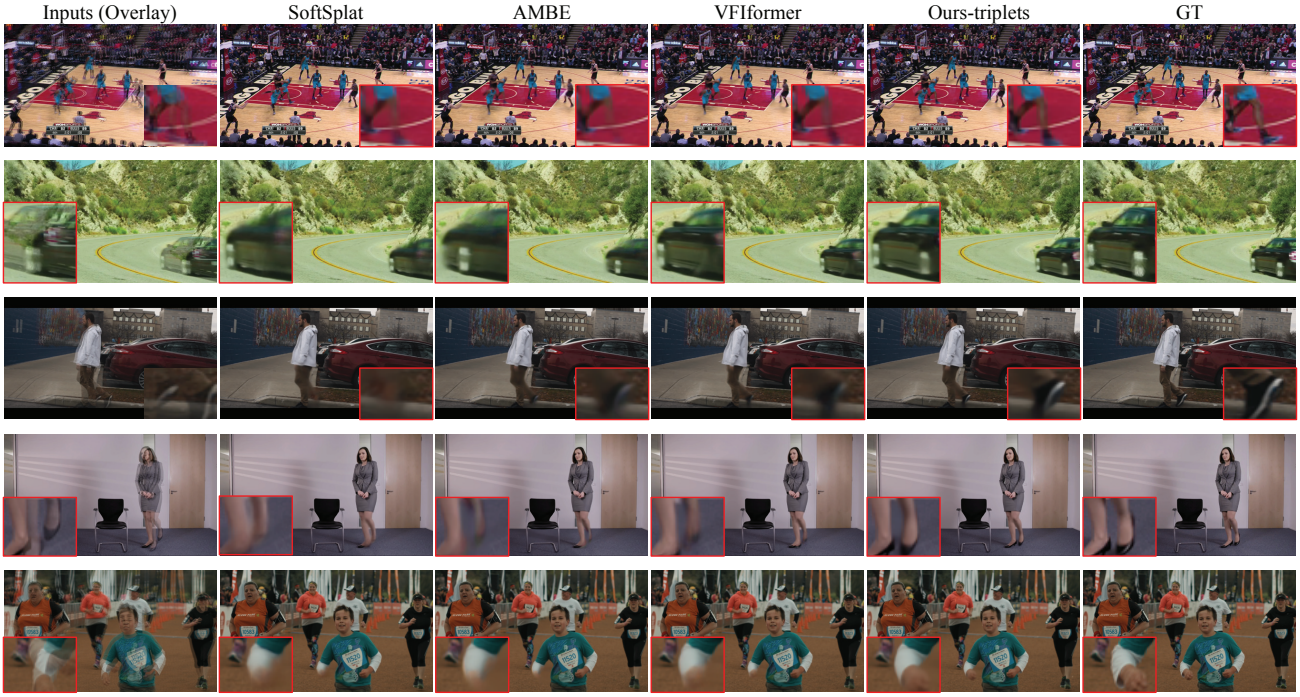
[1] https://media.xiph.org/video/derf/.



(a) Visualized results of the SepConv [7] without/with our TCL on Vimeo-Triplets-Test for video frame interpolation.



(b) Visualized results of our models trained without/with our TCL on Vimeo-Triplets-Test for video frame interpolation



(a) Visual comparison of video frame interpolation on Vimeo-Triplets-Test.

(b) Visual comparison of video frame extrapolation on Vimeo-Triplets-Test.

Figure B.4. Visual comparison of state-of-the-art algorithms. (a-b) refer to the qualitative results of video frame interpolation/extrapolation, respectively. Our method outperforms other state-of-the-art approaches with finer details and fewer artifacts.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[2] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[3] Tarun Kalluri, Deepak Pathak, Manmohan Chandraker, and Du Tran. Flavr: Flow-agnostic video representations for fast frame interpolation. *arXiv preprint arXiv:2012.08512*, 2020. 2

[4] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018. 1

[5] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):346–360, 2013. 1

[6] Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, and Jiaya Jia. Video frame interpolation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3532–3542, 2022. 2

[7] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–270, 2017. 1, 2, 3

[8] Zhihao Shi, Xiangyu Xu, Xiaohong Liu, Jun Chen, and Ming-Hsuan Yang. Video frame interpolation transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17482–17491, 2022. 2

[9] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. Xvfi: extreme video frame interpolation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14489–14498, 2021. 2

[10] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[11] Haochen Zhang, Dong Liu, and Zhiwei Xiong. Is there trade-off between spatial and temporal in video super-resolution? *arXiv preprint arXiv:2003.06141*, 2020. 1