TempFormer: Temporally Consistent Transformer for Video Denoising Supplementary

Mingyang Song[1,2], Yang Zhang[2], and Tunç O. Aydın[2]

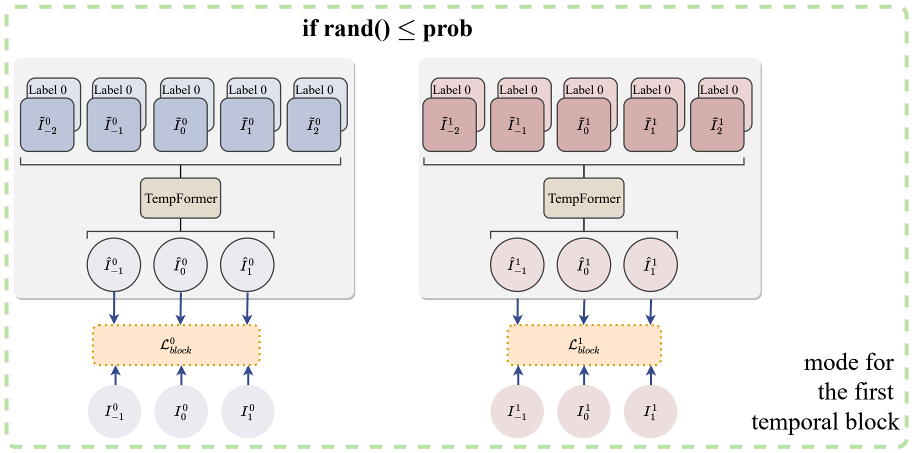
[1] ETH Zurich, Switzerland
[2] DisneyResearch|Studios, Switzerland
`misong@student.ethz.ch`, {`yang.zhang,tunc`}`@disneyresearch.com`

1 Training Strategy

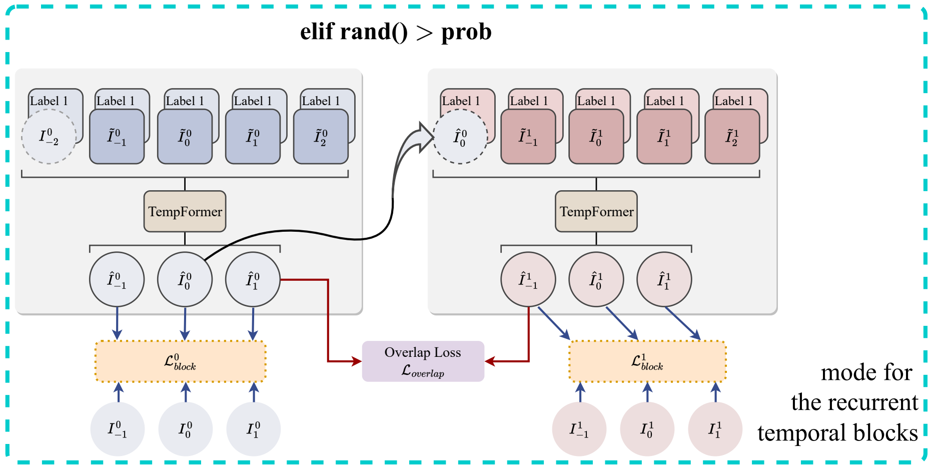
To make our results reproducible, we describe our training strategy in detail in this section. According to our design which has been described in Section 3.2 in the main text, when denoising a video sequence, the inputs of the first temporal block are all noisy, whereas the inputs of each following block contain one denoised frame. The discrepancy in input types may confuse the network and make the model hard to train. To solve this unavoidable problem in the recurrent architecture, Maggioni et al. [2] load multiple frames in each training step to dilute the discrepancy in inputs type. However, the attention layers in our model require large GPU memory during training, which makes it impractical to load more than three blocks in one training step. To solve this problem, we add an extra label to the inputs that indicates their type, i.e., label 0 represents five noisy inputs (mode 0, shown in Figure 1a), and label 1 represents one denoised input followed by four noisy inputs (mode 1, shown in Figure 1b). Furthermore, we observed that when fine-tuning the model always with the recurrent architecture (mode 1), the denoised frames contain checkerboard artifacts if the inputs are all noisy (mode 0). To solve this problem, we generate a random number $rand \sim U(0,1)$ in each training step to choose one mode so that both modes can be trained in Temporal Coherency Enhancement Phase (TCE). In our implementation, the probability that decides which mode to choose is set to 0.5.

2 Testing Details

As described in Section 3.2 in the main text, when testing a noisy video, we split the sequence with a designed stride so that there exists one overlap in the denoised frames of two neighboring blocks. As a result, we have two denoised frames for this position, i.e. $\hat{I}_1^t$ and $\hat{I}_{-1}^{t+1}$. Liang et al. [1] adopted a similar strategy when dividing the sequence and average the overlapping frames to fuse the junctions of neighboring blocks. Such a simple fusion strategy slightly alleviates the altitude of temporal flickering (shown in Figure 2), but the inconsistency artifacts are still obvious in long video sequences. The TCE phase makes the



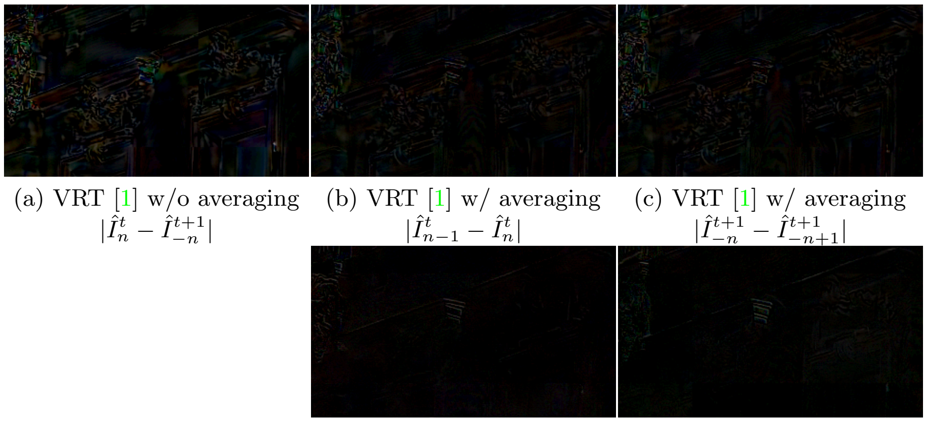
(a) If *rand* ≤ the give probability, the inputs of both blocks are all noisy and labeled with 0.



(b) If *rand* > the give probability, the model is fine-tuned in recurrent architecture and the inputs are labeled with 1.

Fig. 1: Visualization of the training strategy in Temporal Coherency Enhancement Phase.

two overlapping output frames close enough, so choosing the arbitrary denoised frame at this position can produce a temporally stable video. In our implementation, we use the first output from the following block for the overlapping position, namely, $\hat{I}_{-1}^{t+1}$.



(a) VRT [1] w/o averaging $|\hat{I}_{n}^{t} - \hat{I}_{-n}^{t+1}|$

(b) VRT [1] w/ averaging $|\hat{I}_{n-1}^{t} - \hat{I}_{n}^{t}|$

(c) VRT [1] w/ averaging $|\hat{I}_{-n}^{t+1} - \hat{I}_{-n+1}^{t+1}|$

(d) TempFormer $|\hat{I}_{0}^{t} - \hat{I}_{1}^{t}|$

(e) TempFormer $|\hat{I}_{-1}^{t+1} - \hat{I}_{0}^{t+1}|$

Fig. 2: Comparison of the residual images of the denoised frames at the junction of two neighboring blocks. The images are cropped from the reliefs in the video *breakdance* 1080P in DAVIS 2017 [3]. The comparison between (b) and (d) ((c) and (e)) shows that even without averaging on the overlapping frames, with the help of TCE, our model achieves better temporal consistency.

3 Additional Residual Images

Figure 3-4 shows the residual images of the denoised frames by different methods. As described in Section 2, our model did not use any fusion strategy on the overlapping frames. As a result, when comparing with VRT [1] (each temporal block contains twelve frames) and VRT† [1] (each temporal block contains five frames), we change their default configuration so that there are no overlapping frames in their outputs.

4 Additional Temporal Consistency Comparison

Besides the residual images shown in the paper, we provide denoised videos by different methods (contained in the folder ***videos***). To make the difference between each method more evident, we test the sequences under noise level $\sigma = 50$, and crop out the regions that are static or with slight camera movements.

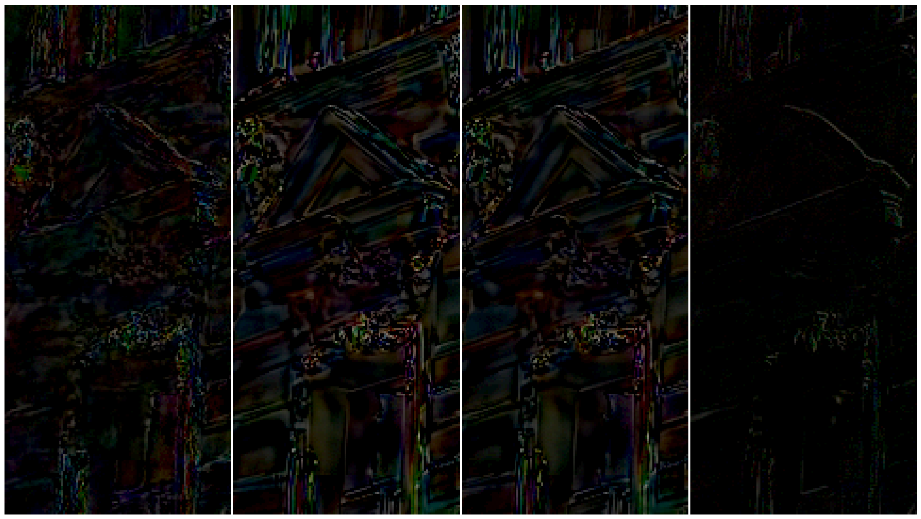


Fig. 3: Residual images of the denoised junction frames from neighboring blocks. The images are cropped from the reliefs in the video *breakdance* 1080P (frame 59 and frame 60) in DAVIS 2017 [3].

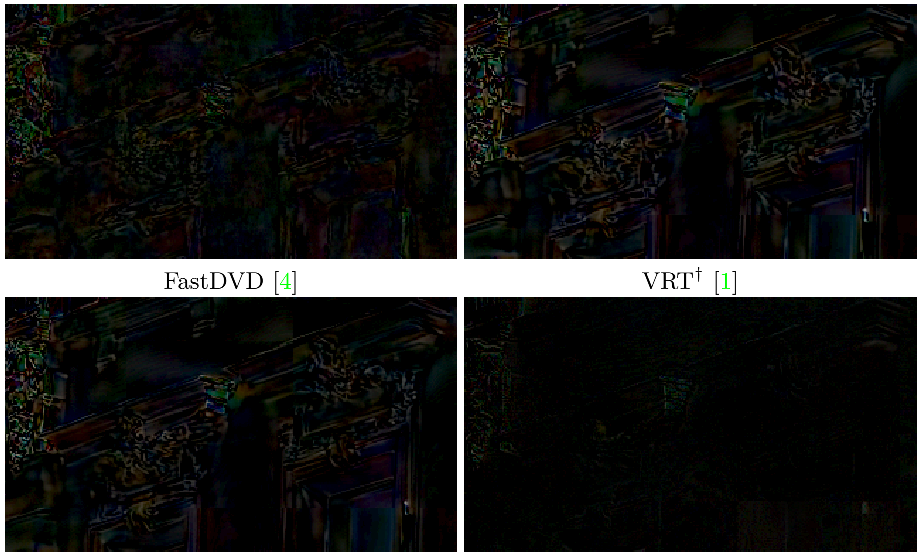


Fig. 4: Residual images of the denoised junction frames from neighboring blocks. The images are cropped from the reliefs in the video *breakdance* 1080P (frame 59 and frame 60) in DAVIS 2017 [3].

All the videos are selected from dataset DAVIS 2017 [3], and the file name shows its source and resolution. Since the videos in DAVIS 2017 [3] are short, which makes some temporal inconsistency artifacts hard to observe, we extend the videos by mirroring the sequence. We compare TempFormer with VRT [1] (each temporal block contains twelve frames), VRT† [1] (each temporal block contains five frames) and FastDVD [4]. Loop the videos for better visualization.

References

1. Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L.: Vrt: A video restoration transformer. arXiv preprint arXiv:2201.12288 (2022) 1, 3, 4, 5
2. Maggioni, M., Huang, Y., Li, C., Xiao, S., Fu, Z., Song, F.: Efficient multi-stage video denoising with recurrent spatio-temporal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3466–3475 (2021) 1
3. 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) 3, 4, 5
4. Tassano, M., Delon, J., Veit, T.: Fastdvdnet: Towards real-time deep video denoising without flow estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1354–1363 (2020) 4, 5