Supplementary Material for
AverNet: All-in-one Video Restoration for
Time-varying Unknown Degradations

Anonymous Author(s)
Affiliation
Address
email

1 Introduction

In this material, we first present more details about degradations in the data synthesis approach and the calculations of parameters and runtime. Then, we present more qualitative results to demonstrate the effectiveness of our AverNet. Finally, we discuss the broader impacts and limitations of this work.

2 Degradations in Video Synthesis

To address data scarcity issue for studying time-varying unknown degradations (TUD), we propose a new approach based on the degradation model in [1, 2] to synthesize corrupted-clean video pairs that contain TUD. In the synthesis pipeline, the video clips are degraded by three major categories of degradations, *i.e.*, noise, blur, and compression. In the following, we will detail the types and parameters of each degradation within these categories.

Noise. In the pipeline, there are three kinds of common noise, *i.e.*, Gaussian noise, Poisson noise, and speckle noise. For Gaussian noise and speckle noise, the noise levels are both uniformly sampled from [10, 15]. The Poisson noise is mathematically modeled as

$$n \sim \mathcal{P}(10^\alpha \times x)/10^\alpha - x, \tag{1}$$

where the α is uniformly sampled from [2,4].

Blur. There are two types of blur in pipeline, *i.e.*, Gaussian blur and resizing blur, which usually appear in action videos and Internet videos. For Gaussian blur, the kernel size is uniformly sampled from {3,5,7}, and the kernel is randomly chosen from {'iso', 'aniso', 'generalized_iso', 'generalized_aniso', 'plateau_iso', 'plateau_aniso'} with the probabilities of {0.405, 0.225, 0.108, 0.027, 0.108, 0.027}. For resizing blur, the resize scale and the interpolation mode is uniformly sampled from [0.5, 2] and {'bilinear', 'area', 'bicubic'}, respectively.

Compression. This degradation includes JPEG and video compression. For JPEG compression, the quality factor is randomly chosen from {20,30,40}. For video compression, the codecs and bitrate are randomly selected from {'libx264', 'h264', ''mpeg4'} and [1e4, 1e5], respectively.

3 Computation of Parameters and Runtime

We compare the parameters and runtime of our AverNet with other methods in the main body of the paper. Specifically, the parameters are computed on a video comprising 24 frames with a resolution of 540×360, while the runtime is computed on a DAVIS-test video consisting of 70 frames. Note that the parameters of EDVR [3] are computed on 5 frames with a 540×360 resolution since it only takes 5 frames as inputs at a time.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

4 Qualitative Results on Videos with TUD

In addition to the qualitative results presented in the main body of the paper, we show more results on the datasets with TUD. To be specific, Fig. 1 and 2 present the qualitative results on DAVIS-test [4] and Set8 [5] datasets with variation intensity $t = 6$. Moreover, Fig. 3 shows the qualitative results in the noise&blur combination on DAVIS-test. As shown in Fig. 1 and 2, BasicVSR++ [6] and Shift-Net [7] leave noise and artifacts in the frames. Furthermore, RVRT [8] produces frames with significant color distortion. In contrast, our method yields results with finer details and less artifacts. From Fig. 3, one could observe that the results of all-in-one image restoration methods are blurry, and video restoration methods BasicVSR++ and RVRT leave artifacts in the results. In contrast, the results of our method have clearer outlines and are closer to the GT.

Figure 1: Qualitative results on the "orchid" video from DAVIS ($t = 6$), from which one could see that existing methods leave noise or artifacts in the results. In contrast, the results of our method have less artifacts and finer details.
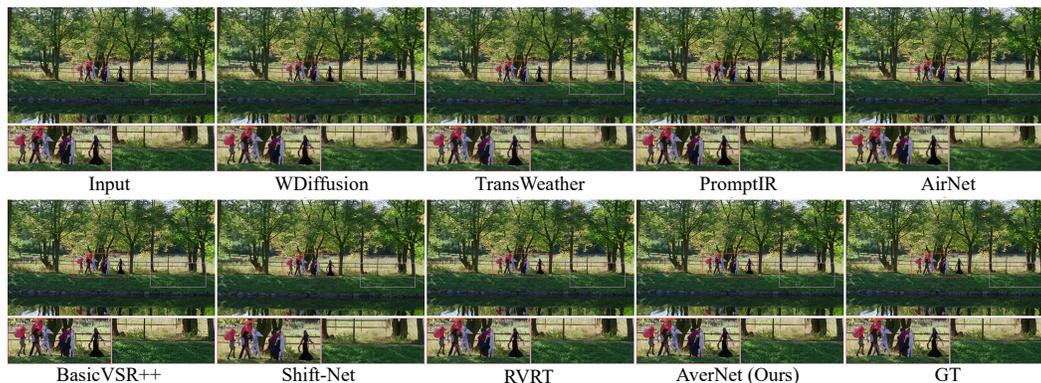
Figure 2: Qualitative results on the "park_joy" video from Set8 ($t = 6$), from which one could observe that existing methods yield blurry or distorted results. In contrast, the results of our methods are clearer and closer to the GT.

5 Broader Impact

In this section, we discuss the impact of our AverNet in a broader vision. Generally, AverNet is the first all-in-one solution to recover videos that contain time-varying unknown degradations, which are prevalent in real-world scenarios. Therefore, it may have multiple applications such as film restoration, surveillance video enhancement, and medical image restoration. However, the videos

2

Figure 3: Qualitative results on the "subway" video from DAVIS-test in the noise&blur degradation combination, from which one could observe that the results of existing methods are blurry. In contrast, the results of our method have clearer outlines and tones that are more similar to the GT.

restored by AverNet may not have the permission of the original copyright holder, thereby infringing the rights of others. Moreover, the training and testing of the model consume a lot of electricity, which causes carbon emissions.

6   Limitation

The training data for AverNet is based on the aforementioned video synthesis approach, which generates videos with time-varying unknown degradations closing to real-world scenarios. However, the corruption in the real-world videos are more complex and hard to be simulated. Therefore, in real-world applications, our AverNet needs further validation and improvement.

References

[1] Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Deng-Ping Fan, Radu Timofte, and Luc Van Gool. Practical blind image denoising via swin-conv-unet and data synthesis. *Machine Intelligence Research*, 20(6):822–836, 2023.

[2] Jiezhang Cao, Qin Wang, Jingyun Liang, Yulun Zhang, Kai Zhang, Radu Timofte, and Luc Van Gool. Learning task-oriented flows to mutually guide feature alignment in synthesized and real video denoising. *arXiv preprint arXiv:2208.11803*, 2022.

[3] 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.

[4] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.

[5] Matias Tassano, Julie Delon, and Thomas Veit. Dvdnet: A fast network for deep video denoising. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1805–1809. IEEE, 2019.

[6] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022.

[7] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9822–9832, 2023.

3

[8] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35:378–393, 2022.

4