

# Supplementary Materials: RainMamba: Enhanced Locality Learning with State Space Models for Video Deraining

Anonymous Authors

In this supplementary material, we present network complexity (Section 1), more ablation studies (Section 2) and extra visual demonstration (Section 3). In addition, a video demo is provided to showcase the dynamic display of our proposed local scanning mechanism and the effectiveness of our method in the supplementary video.

## 1 MODEL ANALYSIS

### 1.1 Model Complexity and Parameters Comparison

As reported in Table 1, we compare the number of parameters, FLOPs, and running time of our network and state-of-the-art methods on a NVIDIA RTX 4090 GPU. The GFLOPs and Runtime are calculated by inferring a video clip of five frames with a resolution of  $256 \times 256$ . We follow [4] to calculate the GFLOPs metric. And the runtime indicates the time needed to process each frame during inference. As shown in Figure 1, we demonstrate the effectiveness of our RainMamba by achieving state-of-the-art results on the VRDS datasets while maintaining a comparatively minimal computational expense. In the presented tabular data, our method obtained the best restoration performance results compared to other comparative methods and achieved the fastest speed. Compared to state-of-the-art method ViMPNet [10], our RainMamba has  $9.51 \times$  fewer FLOPs and runs  $4.96 \times$  faster. Moreover, our model boosts a 4.68 dB improvement in PSNR compared to the second fastest method ESTINet [17], and achieving an inference speed of 89.2 FPS. This relatively fast inference speed enables our model to be effectively utilized in real-world applications. Thanks to the linear complexity of state space models and the critical components of our network, our approach achieves significant improvements in both performance and speed.

## 2 MORE ABLATION STUDIES

### 2.1 Visual Results of Ablation Study

As shown in Figure 2, in addition to quantitatively comparing the ablation experiments of RainMamba on VRDS dataset, we also conducted visual comparisons to qualitatively verify the effectiveness of three critical components of our network. By introducing the Global Mamba Block (GMB), our “M2” model effectively eliminates a significant number of artifacts associated with raindrops and rain streaks, compared to “M1”. However, the “M3” model can not preserve the spatial structure of certain details effectively and introduced extensive artifacts outside the window. Leveraging Local Mamba Block (LMB), our “M3” model achieves superior detail retention, such as the shape of windows. The integration of GMB and LMB significantly enhances the detail recovery in areas obscured by raindrops, as our “M4” model improves the modeling ability of spatiotemporal information. By combining these three complementary contributions, our RainMamba clearly removes the artifacts and better recovers the scene structures.

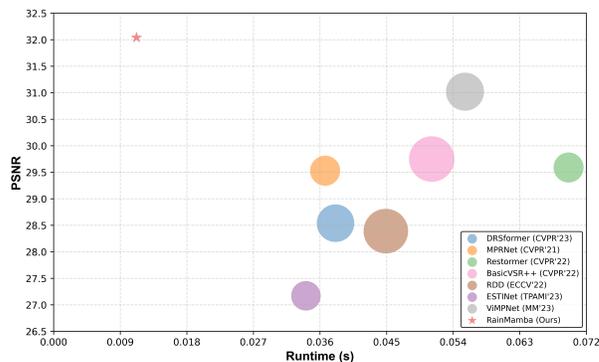


Figure 1: PSNR performance v.s Runtime and GFLOPs on VRDS dataset. The size of the circles and pentagram indicates the GFLOPs of model.

Table 1: Model complexity and parameters comparisons between our network and other methods. Bolded and underlined values indicate the best and the second-best performance, respectively.

Method	PSNR	SSIM	LPIPS	GFLOPs	Runtime(s/frame)	Parameters(M)
DRSformer[3]	28.54	0.9075	0.1143	1101.89	0.0381	33.63
MPRNet[16]	29.53	0.9175	0.0987	706.19	0.0367	<b>3.64</b>
Restormer[15]	29.59	0.9206	0.0925	<u>704.95</u>	0.0696	26.10M
BasicVSR++[1]	29.75	0.9171	0.1023	1616.44	0.0511	6.22
RDD[8]	28.39	0.9096	0.1168	1553.65	0.0449	<u>5.53</u>
ESTINet [17]	27.17	0.8436	0.2253	681.83	<u>0.0341</u>	22.96
ViMPNet [10]	<u>31.02</u>	<u>0.9283</u>	<u>0.0862</u>	1131.7	0.0556	32.10
Ours	<b>32.04</b>	<b>0.9366</b>	<b>0.0684</b>	<b>118.99</b>	<b>0.0112</b>	34.75

Table 2: Quantitative comparisons between our network and SOTA methods on the NTURain dataset [2].

Methods	PSNR	SSIM
MSCSC [5]	27.31	0.7870
J4RNet [6]	32.14	0.9480
SPAC [2]	33.11	0.9474
FCRNet [11]	36.05	0.9676
SLDNet[13]	34.89	0.9540
MPRNet[16]	36.11	0.9637
S2VD [14]	37.37	0.9683
ESTINet [17]	37.48	0.9700
Ours	<b>37.87</b>	<b>0.9738</b>

### 2.2 Visual Comparisons of Different Scanning Mechanisms

Figure 3 illustrates the results of “M2” and “M3” model. The “M2” model can generate incorrect directional extensions when reconstructing the shapes of objects obscured by raindrops. This issue arises because “M2” utilizes a global scanning approach (row-and-column-major order), which neglects spatio-temporal continuity and leads to local pixel forgetting. Our proposed local scanning

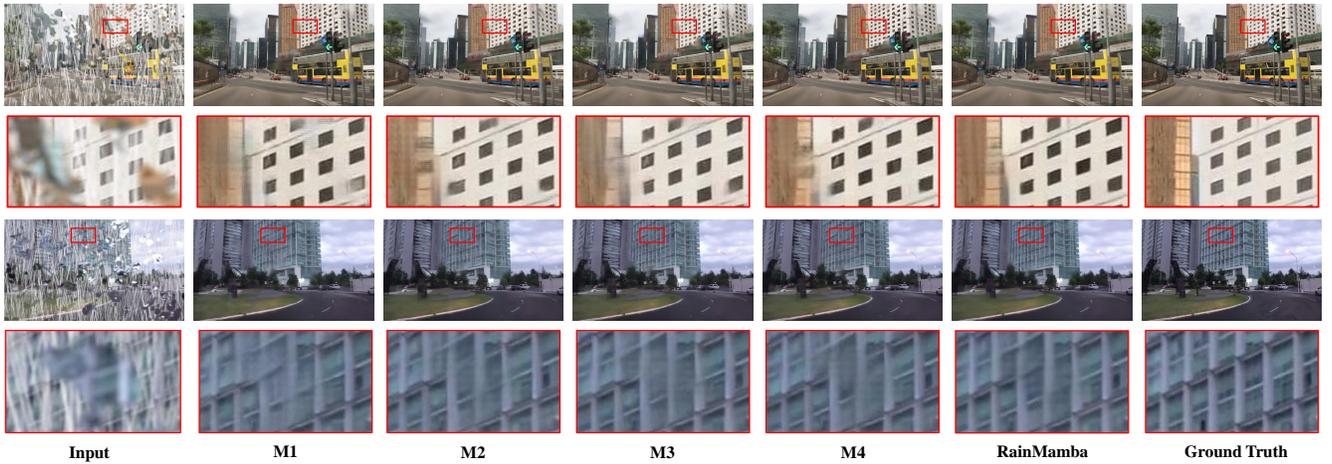


Figure 2: Visual comparisons of the ablation study on input video frames from the VRDS dataset. (Please zoom in for a better illustration.)

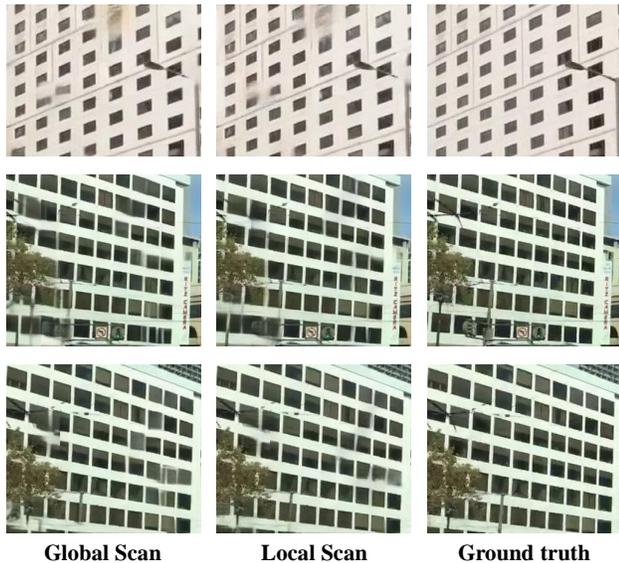


Figure 3: Visual comparisons of two different scanning mechanisms. Our local scanning mechanism improves spatial structure preservation of derained results.

mechanism improves the network’s local information awareness by rearranging the scan path of Mamba in sequence-level temporal modeling.

### 3 MORE EXPERIMENTAL RESULTS

#### 3.1 More Implementation Details

The encoder extract multi-scale feature maps (i.e., scales of  $1/4$ ,  $1/8$ ,  $1/16$ ,  $1/32$ ), after which a lightweight head merges these maps to generate encoded features  $E_t$ . We set the hyperparameters for different datasets according to the original paper’s settings. For the RainVID&SS[7] and RainSynAll100[12] datasets, input frames

Table 3: Analysis of long video processing by our network on the NTURain dataset [2].

Frames	PSNR	SSIM	Memory
7	37.534	0.96904	5,082M
10	37.722	0.97331	6,026M
20	37.806	0.97357	9,530M
30	37.838	0.97367	13,054M
40	37.847	0.97371	17,296M
50	37.858	0.97374	20,056M
60	37.863	0.97376	23,556M
70	37.867	0.97378	27,068M
80	37.870	0.97379	32,056M
90	37.875	0.97380	34,090M
100	37.876	0.97380	37,594M
110	37.875	0.97380	41,100M

are randomly cropped to a spatial resolution of  $128 \times 128$ , with the number of frames per video clip being 7 and 5 respectively. For the LWDDS dataset [9], the input frame is cropped to  $256 \times 256$ , with 5 frames per clip. The initial learning rate for our network is set at  $2 \times 10^{-4}$  for RainVID&SS and RainSynAll100 datasets, and  $4 \times 10^{-4}$  for the LWDDS dataset. A consistent batch size of 4 is used across these three datasets.

#### 3.2 Quantitative and Qualitative Comparisons on the NTURain Dataset

We also conducted comparisons of our model with state-of-the-art video deraining methods on a widely-utilized NTURain dataset for video rain streak removal. NTURain [2] dataset contains 25 videos for training and 8 videos for testing. From these quantitative results in Tab. 2, Our method demonstrates an enhancement in performance over the ESTINet [17], improving the PSNR score from 37.48 dB to 37.87 dB, and the SSIM score from 0.9700 to 0.9738. These results demonstrate the capability of our method to effectively

remove rain streaks in videos without incorporating any additional physical priors.

Figure 4 visually compares rain streak removal results predicted by our network and state-of-the-art method ESTINet [17] from the NTURain dataset. Compared with ESTINet, our network demonstrates superior performance in restoring the original background images by effectively eliminating rain streaks from input video frames.

### 3.3 Analysis of Long Video Processing

We selected the NTURain dataset as our test dataset due to its lower frame resolution (640×480) and the extensive sequence length of its videos (ranging from 116 to 298 frames). Our experiment is implemented on a NVIDIA RTX A6000 GPU with a graphics memory of 48 GB. We initially select 7 frames for input according to the training setting, and subsequently increased the number of input frames in increments of 10. As reported in Table 3, we input full resolution video clips and compare the experimental results from using video clips of different lengths. The experimental results indicate that as the input frame rate increases, the effectiveness of video restoration also improves. This demonstrates that the long-sequence modeling capability of SSMs can effectively leverage the spatio-temporal contextual information in videos to successfully remove rain streaks. It is noteworthy that our RainMamba is capable of processing 110 frames of full-resolution video simultaneously on a single GPU. Due to the critical role of inter-frame information in video restoration tasks, the long video processing capabilities of our RainMamba are anticipated to be applicable to other video restoration challenges.

### 3.4 More Results on the Compared Datasets

Figure 5 and Figure 6 demonstrate more visual comparisons between the results generated by our methods and the other compared methods on the RainSynAll100 dataset and the LWDDS dataset, respectively. The results from two datasets show that our RainMamba effectively removes various rain patterns, including streaks and raindrops of different sizes. Also, our approach preserves the most natural color compared to alternative comparative methods.

291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348

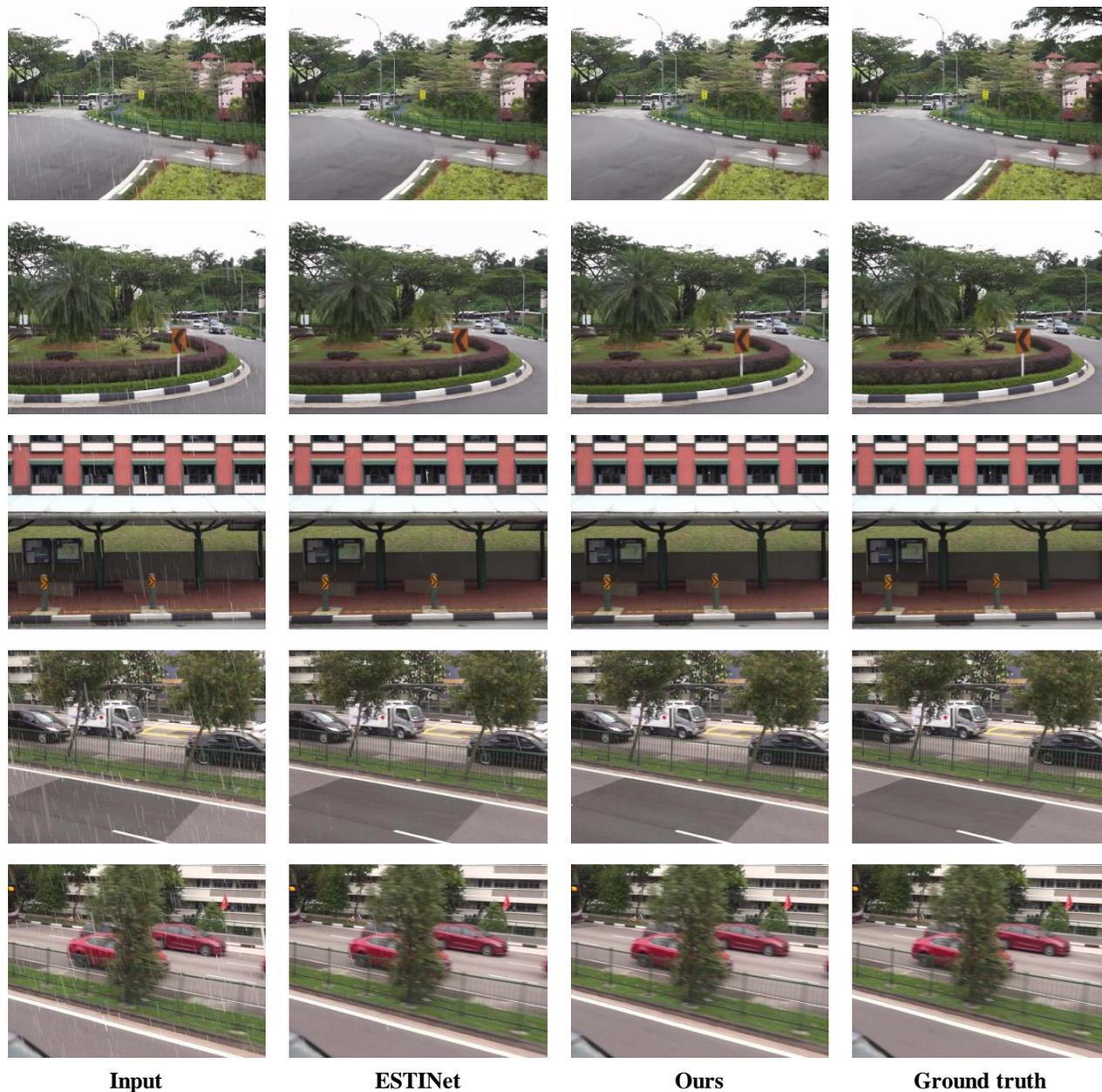


Figure 4: Visual comparisons of derained results from our network and ESTINet [17] on input video frames from the NTURain dataset. (Please zoom in for a better illustration.)

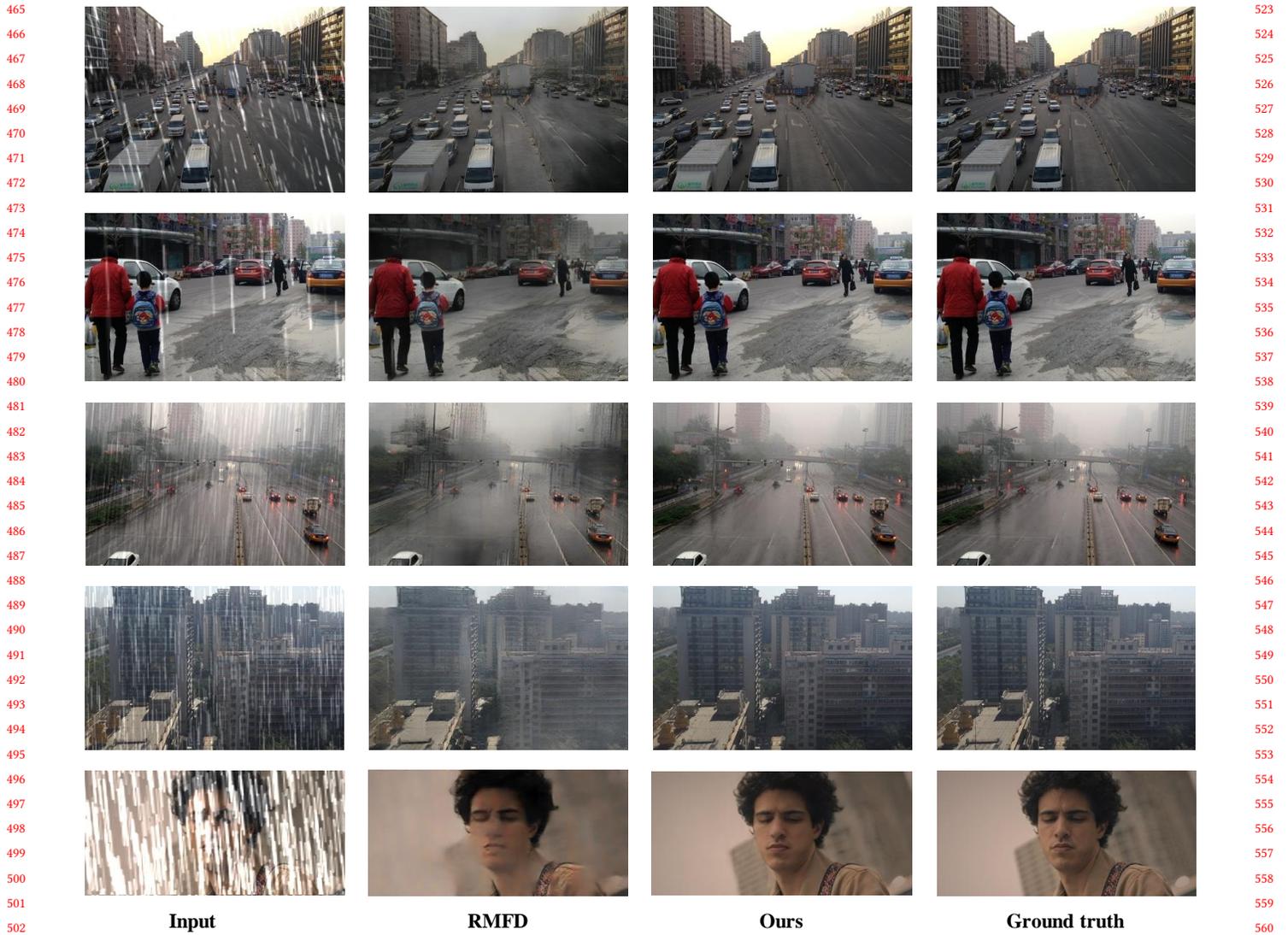


Figure 5: Visual comparisons of derained results from our network and RMFD [12] on input video frames from the RainSynAll100 dataset. (Please zoom in for a better illustration.)

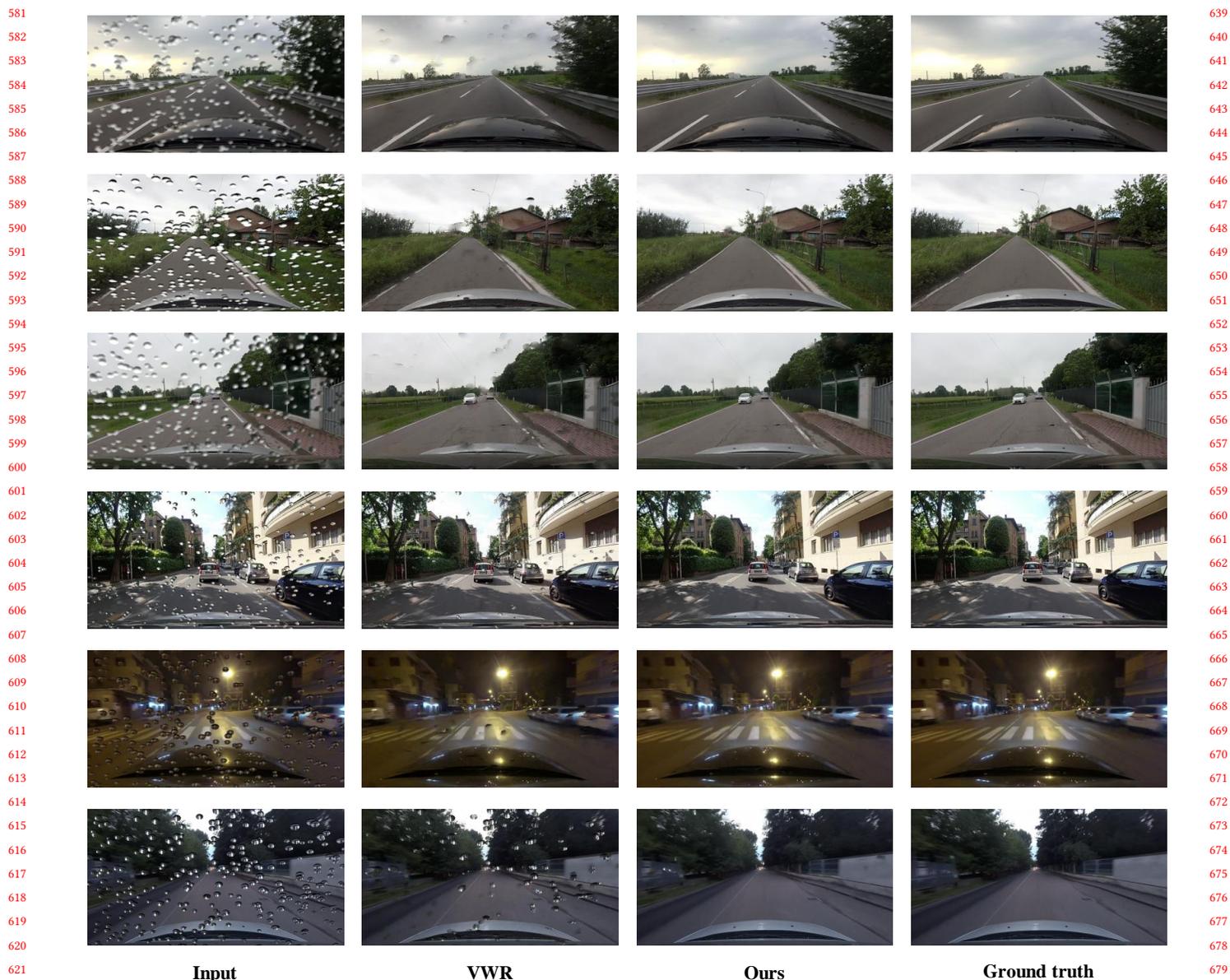


Figure 6: Visual comparisons of derained results from our network and VWR [9] on input video frames from the LWDDS dataset. (Please zoom in for a better illustration.)

## REFERENCES

- [1] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. 2022. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5972–5981.
- [2] Jie Chen, Cheen-Hau Tan, Junhui Hou, Lap-Pui Chau, and He Li. 2018. Robust video content alignment and compensation for rain removal in a cnn framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6286–6295.
- [3] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. 2023. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5896–5905.
- [4] Kunchang Li, Xinhao Li, Yi Wang, Yanan He, Yali Wang, Limin Wang, and Yu Qiao. 2024. VideoMamba: State Space Model for Efficient Video Understanding. *arXiv preprint arXiv:2403.06977* (2024).
- [5] Minghan Li, Qi Xie, Qian Zhao, Wei Wei, Shuhang Gu, Jing Tao, and Deyu Meng. 2018. Video rain streak removal by multiscale convolutional sparse coding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6644–6653.
- [6] Jiaying Liu, Wenhan Yang, Shuai Yang, and Zongming Guo. 2018. Erase or fill? deep joint recurrent rain removal and reconstruction in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3233–3242.
- [7] Shangquan Sun, Wenqi Ren, Jingzhi Li, Kaihao Zhang, Meiyu Liang, and Xiaochun Cao. 2023. Event-aware video deraining via multi-patch progressive learning. *IEEE Transactions on Image Processing* (2023).
- [8] Shuai Wang, Lei Zhu, Huazhu Fu, Jing Qin, Carola-Bibiane Schönlieb, Wei Feng, and Song Wang. 2022. Rethinking Video Rain Streak Removal: A New Synthesis Model and a Deraining Network with Video Rain Prior. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*. Springer, 565–582.
- [9] Qiang Wen, Yue Wu, and Qifeng Chen. 2023. Video Waterdrop Removal via Spatio-Temporal Fusion in Driving Scenes. *arXiv preprint arXiv:2302.05916* (2023).
- [10] Hongtao Wu, Yijun Yang, Haoyu Chen, Jingjing Ren, and Lei Zhu. 2023. Mask-Guided Progressive Network for Joint Raindrop and Rain Streak Removal in Videos. In *Proceedings of the 31st ACM International Conference on Multimedia*. 7216–7225.
- [11] Wenhan Yang, Jiaying Liu, and Jiashi Feng. 2019. Frame-consistent recurrent video deraining with dual-level flow. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1661–1670.
- [12] Wenhan Yang, Robby T Tan, Jiashi Feng, Shiqi Wang, Bin Cheng, and Jiaying Liu. 2021. Recurrent multi-frame deraining: Combining physics guidance and adversarial learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 8569–8586.
- [13] Wenhan Yang, Robby T Tan, Shiqi Wang, and Jiaying Liu. 2020. Self-learning video rain streak removal: When cyclic consistency meets temporal correspondence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1720–1729.
- [14] Zongsheng Yue, Jianwen Xie, Qian Zhao, and Deyu Meng. 2021. Semi-supervised video deraining with dynamical rain generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 642–652.
- [15] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5728–5739.
- [16] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. 2021. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14821–14831.
- [17] Kaihao Zhang, Dongxu Li, Wenhan Luo, Wenqi Ren, and Wei Liu. 2022. Enhanced spatio-temporal interaction learning for video deraining: faster and better. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 1287–1293.