Blur-aware Spatio-temporal Sparse Transformer for Video Deblurring Supplementary Material

In this supplementary material, we offer more details and additional results to complement the main paper. First, we offer more details of network architecture in Sec. A. Next, we present additional ablation study results in Sec. B. Finally, more qualitative comparisons are shown in Sec. C.

A. Architecture Details

Encoder and Decoder. The BSSTNet employs the same Channel Attention Block (CAB) and Grouped Spatial-temporal Shift (GSTS) in the encoder and decoder, following Shift-Net+ [1].

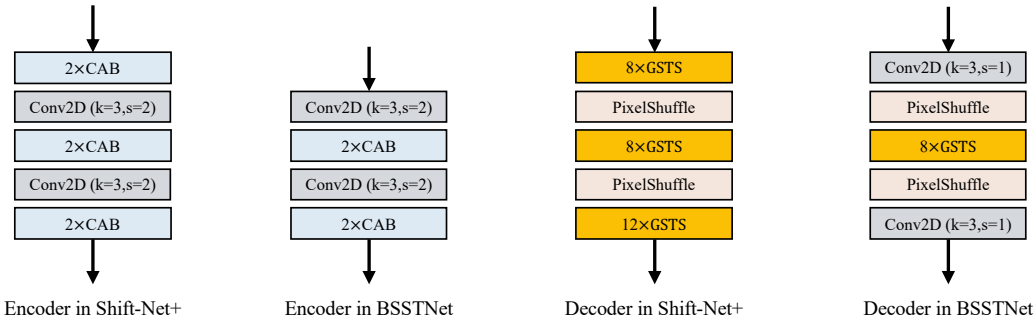


Figure 1. **Encoder-Decoder structure in BSSTNet.** k and s represent the convolution kernel size and stride, respectively.

As shown in Figure 1, considering the computational complexity, BSSTNet employs a smaller Encoder-Decoder compared to Shift-Net+.

B. Additional Ablation Study Results

Spatial Sparsity Levels in BSST. In the main paper, we fix the value of θ at 0.3. Different choices of θ lead to varying levels of spatial sparsity in BSST. Table 1 illustrates different θ configurations and their impact on both computational complexity and performance. The results indicate that going beyond $\theta = 0.3$ adversely affects performance, and there is no significant performance gain when θ is less than this threshold. To maintain a balance between performance and computational complexity, we choose $\theta = 0.3$.

Table 1. **Comparison between different θ configurations and their impact on both computational complexity and performance.** Note that the results are evaluated on the DVD dataset. The percentage symbol (%) represents the proportion of selected tokens in the spatial domain of BSST.

θ	PSNR	SSIM	GFLOPs	%
0.1	35.01	0.9706	142	98
0.2	34.98	0.9704	138	81
0.3	34.95	0.9703	133	42
0.4	34.82	0.9695	128	24
0.5	34.54	0.9682	121	10

The Impact of Different Flow Estimators. We replace RAFT [5] with SPyNet [4]. As shown in Table 2, employing RAFT as the flow estimator in BSSTNet leads to a **0.07 dB** improvement in PSNR and a **0.0005** increase in SSIM compared to using SPyNet, accompanied by a rise in computational complexity by **3 GFLOPs**. The results indicate that a flow estimator with better robustness for blurry frames can lead a slight performance improvement for BSSTNet.

Table 2. **Comparison between different flow estimators.** The best results are highlighted in bold. Note that the results are evaluated on the DVD dataset.

Flow Estimator	PSNR	SSIM	GFLOPs
SPyNet [4]	34.88	0.9698	**130**
RAFT [5]	**34.95**	**0.9703**	133

Effectiveness of Blur Maps. In BSSTNet, the blur maps are generated from the optical flows estimated from blurry frames. Table 3 shows a comparison between the blur maps estimated from blurry frames and sharp frames. As shown in Table 3, the performance of blur maps from blurry frames is comparable to that of blur maps from sharp frames. The results suggest that deriving the blur map from the optical flows of blurry frames does not notably impact performance. Visual results of blur maps and their corresponding blurry frames on the GoPro and DVD datasets are presented in Figures 2, 3, 4, 5, 6, and 7, respectively.

Table 3. **Comparison between the blur maps computed from the optical flows of sharp frames and those computed from the optical flows of blurry frames.** The best results are highlighted in bold. Note that the results are evaluated on the DVD dataset.

	PSNR	SSIM
Blur map from sharp frames	**35.01**	**0.9706**
Blur map from blurry frames	34.95	0.9703

Figure 2. **Visual results of blur maps and corresponding blurry frames on GoPro dataset.** Specifically, we provide the visual results for a sequence length of 8 frames. In the blur maps, a higher weight indicates that the corresponding region in blurry frame is more blurred. The visual results demonstrate that our blur maps accurately provide the information about blurry regions in blurry frames.

Figure 3. **Visual results of blur maps and corresponding blurry frames on GoPro dataset.** Specifically, we provide the visual results for a sequence length of 8 frames. In the blur maps, a higher weight indicates that the corresponding region in blurry frame is more blurred. The visual results demonstrate that our blur maps accurately provide the information about blurry regions in blurry frames.

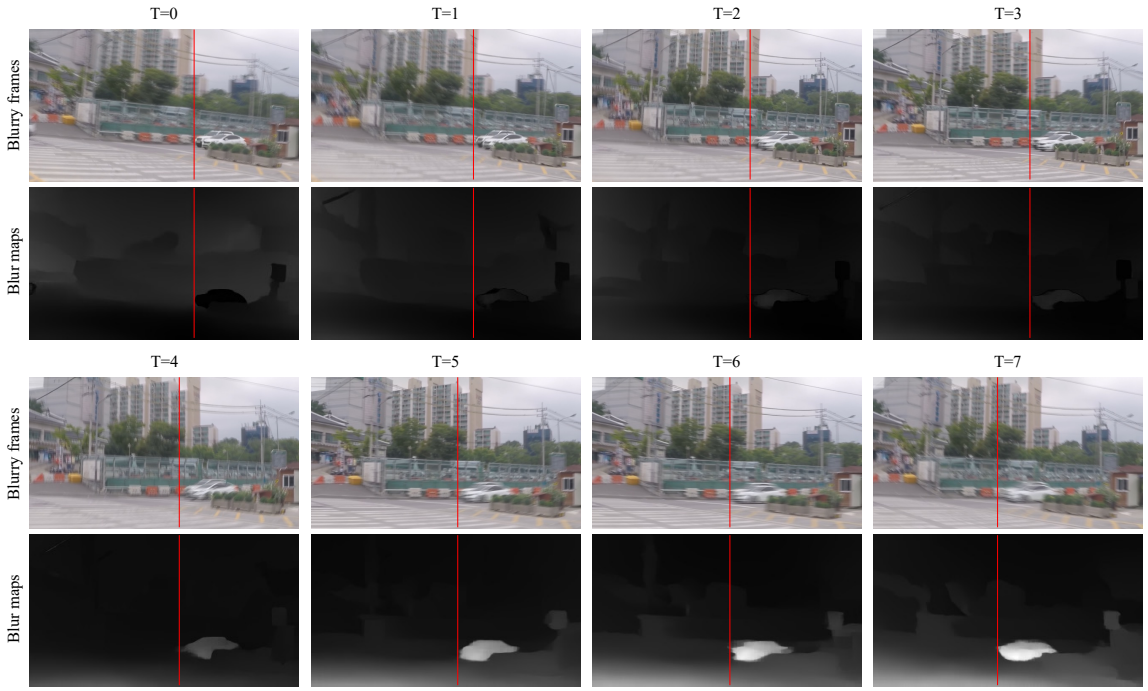


Figure 4. **Visual results of blur maps and corresponding blurry frames on GoPro dataset.** Specifically, we provide the visual results for a sequence length of 8 frames. In the blur maps, a higher weight indicates that the corresponding region in blurry frame is more blurred. The visual results demonstrate that our blur maps accurately provide the information about blurry regions in blurry frames.

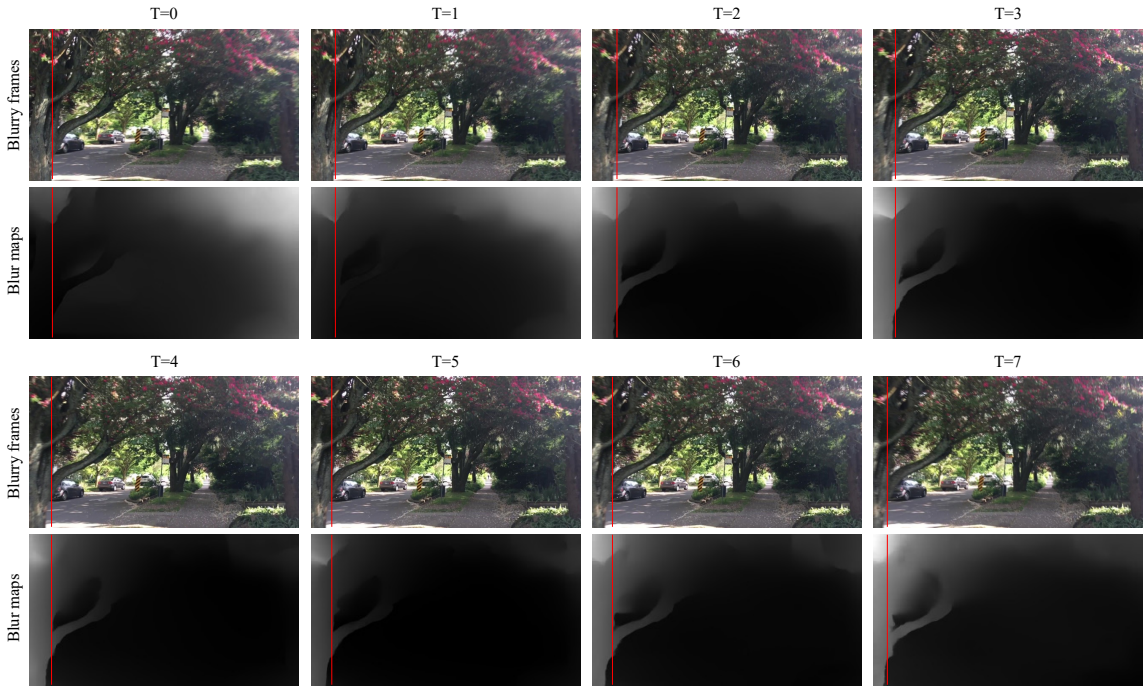


Figure 5. **Visual results of blur maps and corresponding blurry frames on DVD dataset.** Specifically, we provide the visual results for a sequence length of 8 frames. In the blur maps, a higher weight indicates that the corresponding region in blurry frame is more blurred. The visual results demonstrate that our blur maps accurately provide the information about blurry regions in blurry frames.

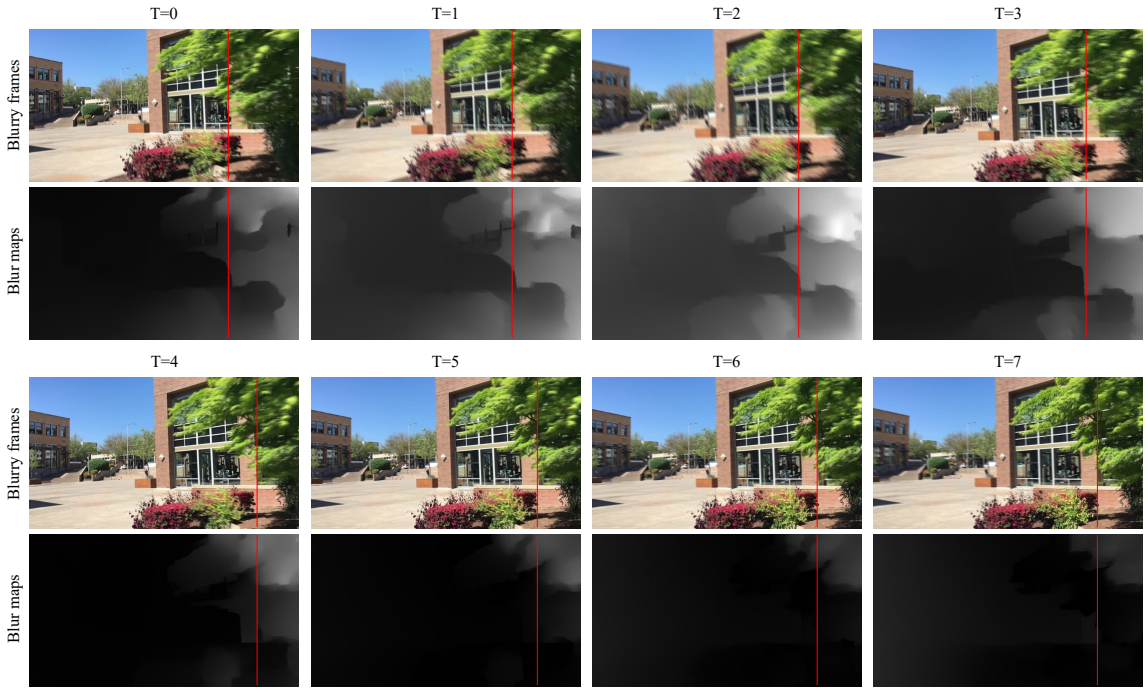


Figure 6. **Visual results of blur maps and corresponding blurry frames on DVD dataset.** Specifically, we provide the visual results for a sequence length of 8 frames. In the blur maps, a higher weight indicates that the corresponding region in blurry frame is more blurred. The visual results demonstrate that our blur maps accurately provide the information about blurry regions in blurry frames.

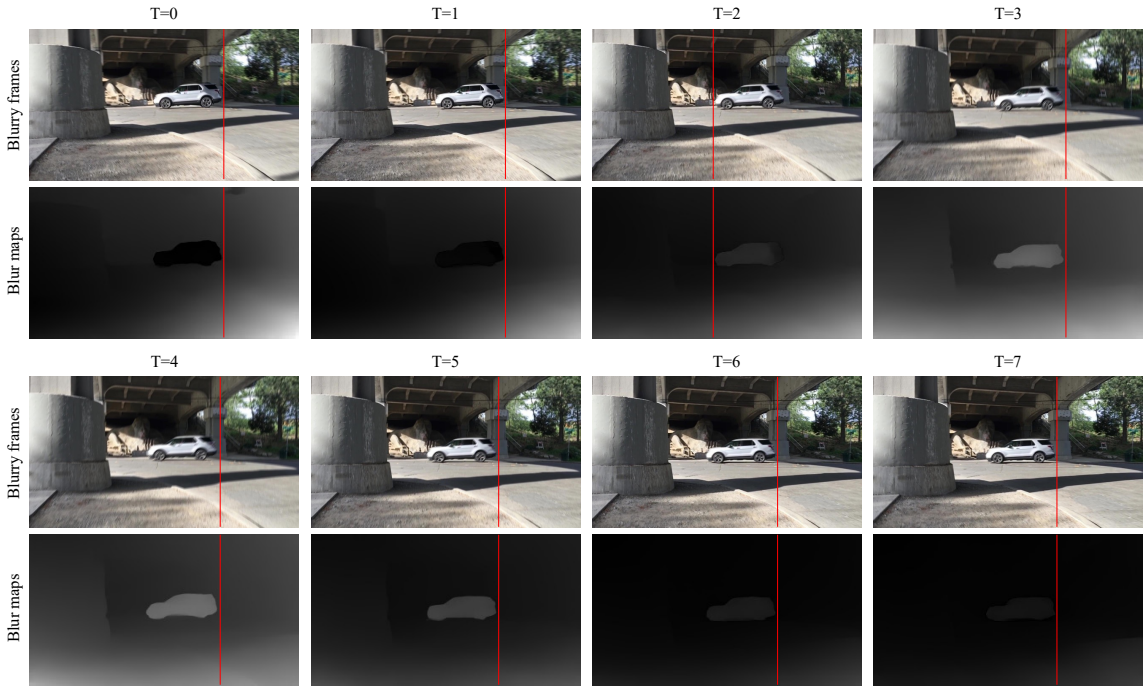
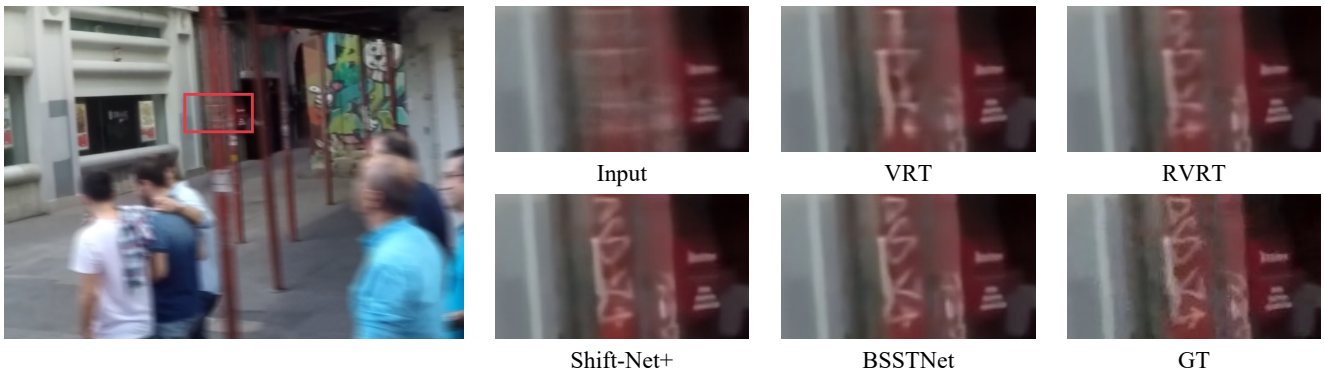


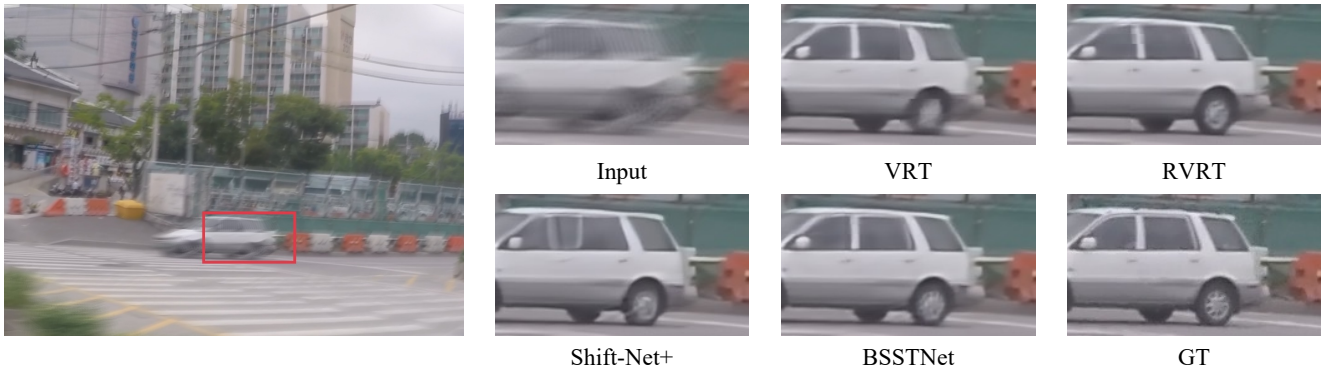
Figure 7. **Visual results of blur maps and corresponding blurry frames on DVD dataset.** Specifically, we provide the visual results for a sequence length of 8 frames. In the blur maps, a higher weight indicates that the corresponding region in blurry frame is more blurred. The visual results demonstrate that our blur maps accurately provide the information about blurry regions in blurry frames.

C. Additional Qualitative Comparisons

C.1. Qualitative Comparisons on GoPro and DVD

In this section, we provide additional qualitative comparisons of BSSTNet with the state-of-the-art methods, including VRT [2], RVRT [3], and Shift-Net+ [1]. Figures 8 and 9 are the qualitative comparisons on the GoPro and DVD datasets, respectively.





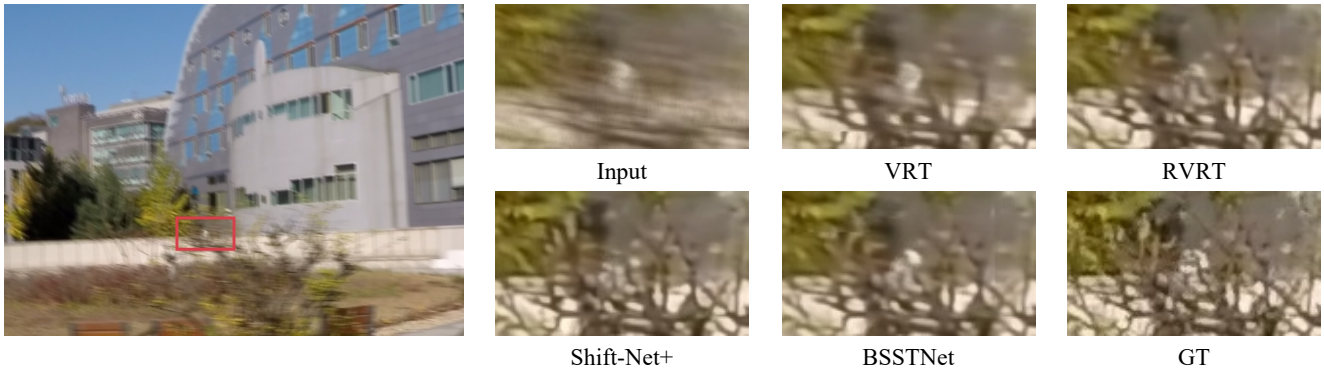


Figure 8. **Qualitative comparison on the GoPro dataset.** Note that "GT" stands for "Ground Truth". The proposed BSSTNet produces images with enhanced sharpness and more detailed visuals compared to competing methods.

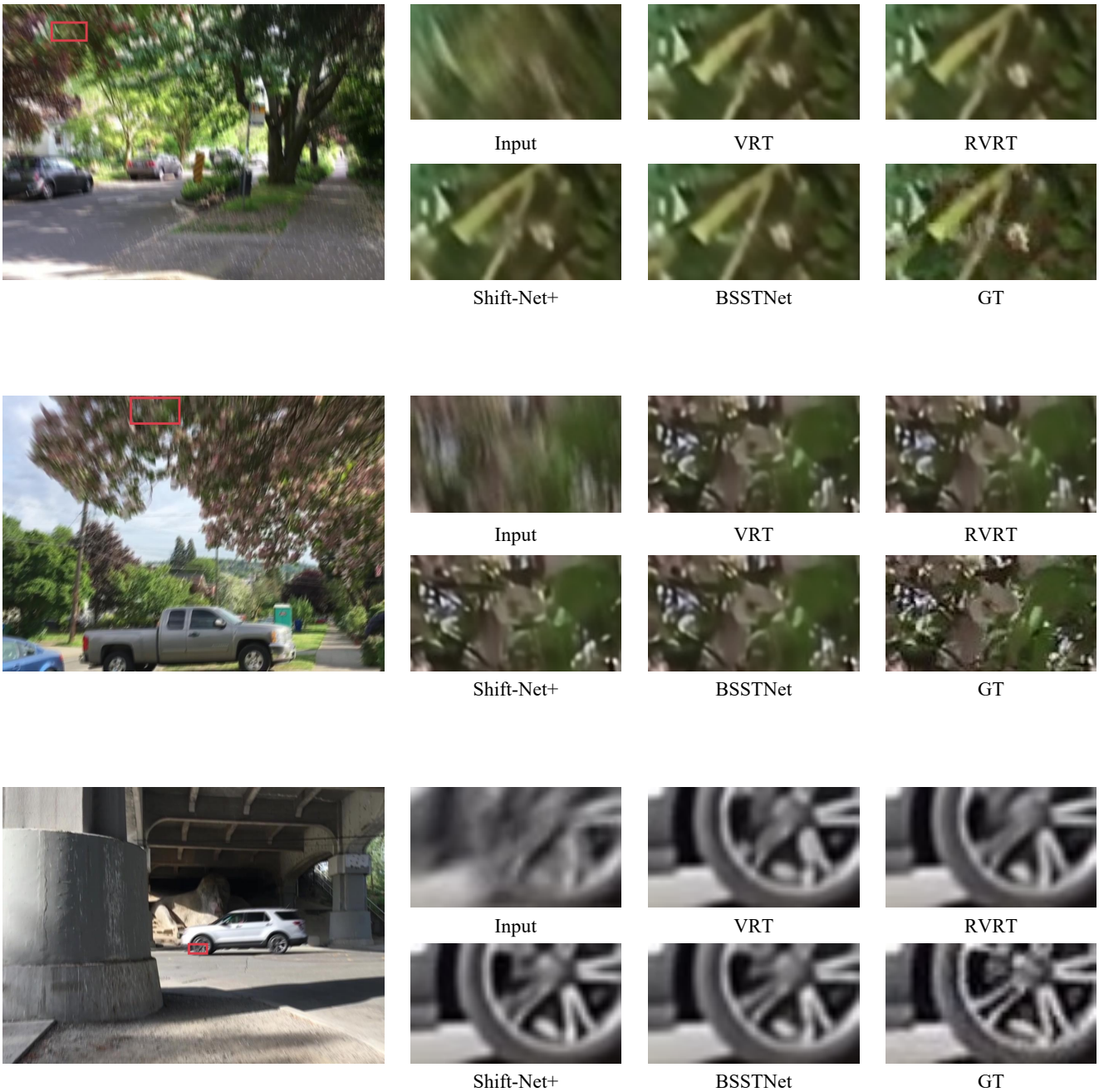


Figure 9. **Qualitative comparison on the DVD dataset.** Note that "GT" stands for "Ground Truth". The proposed BSSTNet produces images with enhanced sharpness and more detailed visuals compared to competing methods.

C.2. Qualitative Comparisons on Real Blurry Videos

Additionally, we present qualitative comparisons of BSSTNet with state-of-the-art methods on real blurry videos. The results can be observed in the https://vilab.hit.edu.cn/projects/bsstnet/.

References

[1] 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 *CVPR*, 2023. 1, 6

[2] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. VRT: A video restoration transformer. arXiv: 2201.12288, 2022. 6

[3] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. In *NeurIPS*, 2022. 6

[4] Anurag Ranjan and Michael J. Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, 2017. 2

[5] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 2