Rethinking Video Frame Interpolation from Shutter Mode Induced Degradation —Supplemental Material—

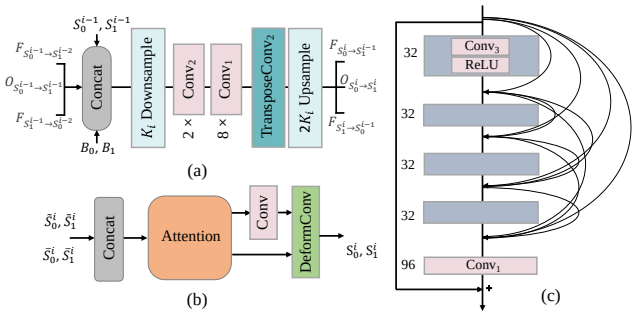
Xiang Ji[1] Zhixiang Wang[1,2] Zhihang Zhong [1,2] Yinqiang Zheng[1†]
[1]The University of Tokyo, Japan [2]National Institute of Informatics, Japan
{jixiang,wangzhixiang}@g.ecc.u-tokyo.ac.jp,
zhong@is.s.u-tokyo.ac.jp, yqzheng@ai.u-tokyo.ac.jp

In this supplemental material, we present implementation details (Section A) and additional results (Section B) as complements of the main content.

A. Implementation Details

A.1. Architecture Design

To explore a generic and adaptive model that could interpolate frames from different degradations, we revisit recent state of the art methods for RS and blur video frame interpolation. By incorporating the advantages of two paradigms for RS correction and motion deblurring, we propose our PMB-Net that decouples the task into correction and interpolation branches with a mutual boosting manner.



Extended Figure A.1: **Architecture of some modules in PMB-Net.** (a) Implementation details of BFP module. (b) Deformable attention layer. (c) Residual dense block.

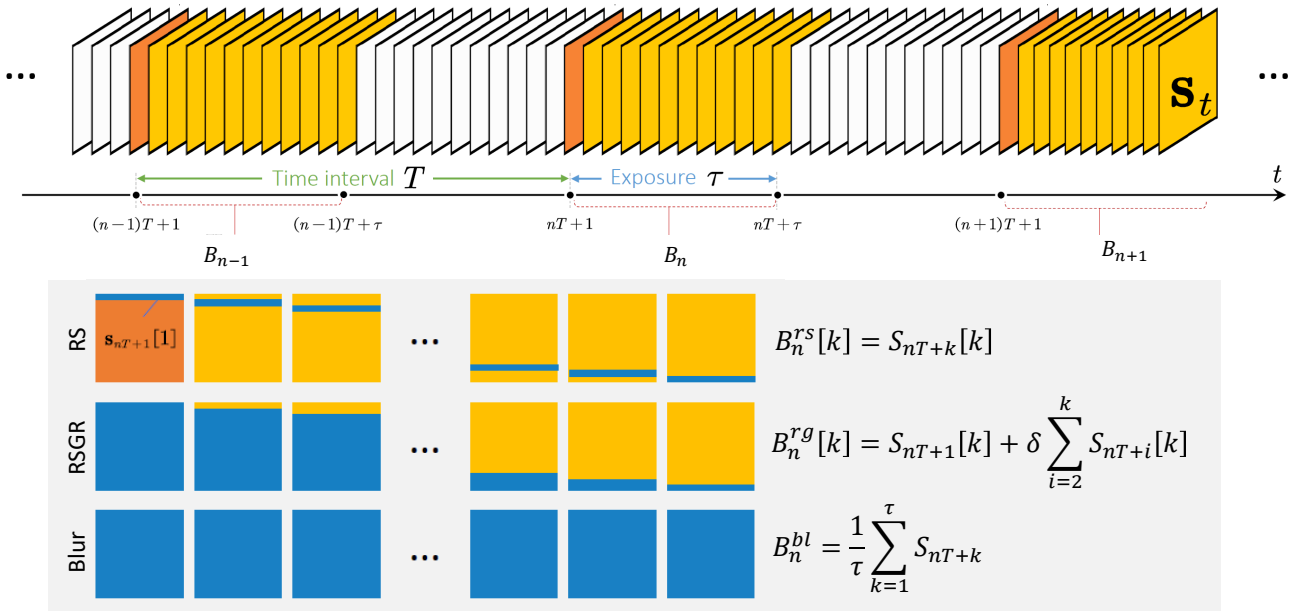
Beside the core parts demonstrated in the main manuscript, we present the more architecture details of implementation. Following the classical bidirectional flow estimation, we construct our BFP module as shown in Fig. A.1(a). Similar to [1], we exploit the SOTA structure of nonwarping deblurring, combined with the multi-input and multi-output strategy to implement our NWD module that can handle multi-scale blur with low computation loads. Asymmetric feature fusion (AFF) is also used to promote propagation of information flow between encoders and decoders. The deformable attention layer in Fig. A.1(b) is exploited to fuse the corrected frames from two streams and we use residual dense layer [4] as our building block in contextual module, which is shown in Fig. A.1(c).

Extended Figure A.2: **Data samples of different degradations under the same scene from our RD-VFI**. On the left, original images are on the top row and corresponding difference maps computed with the HS frame are at bottom.

A.2. Construction of Imaging system

Inspired by recent co-axis optical settings of [9, 11], we construct a quad-axis imaging system to collect realistic dataset of strictly aligned RS, RSGR, GS and high speed videos. We first fix the RS camera, and adjust the orientation

†Corresponding author



Extended Figure A.3: **Synthetic method**. The notation $B[k]$ denotes extracting the k-th row from frame B

and position of the other three cameras by examining the residual images of a checker pattern. We pay special attention to the first scan-line of the RS camera and the RSGR camera, and make sure they are rigorously aligned. This is to preclude motion-related misalignment caused by the time delay between pixel rows of the RS camera (and RSGR camera). However, to achieve this accurate alignment for all four cameras in all directions is extremely difficult, due to complexity of the quad-axis system. We then fix all components and calibrate the alignment of four cameras by using three homographies, of which the close-loop constraint is considered. Since the three beamsplitters will reduce incoming light to a quarter, we capture our dataset in a sunny day so that all images are bright enough.

As for the HS camera, BTRAN CS-700C, it is an air-forced cooling camera with a SONY IMX426 sensor inside. The noise level is very low because of the following two reasons. First, IMX426 sensor has a large pitch size of 9um $\times$ 9um, by combining four 4.5um $\times$ 4.5um subpixels at the circuit level. This greatly enhances its sensitivity and reduces its noise level. Second, we cool the sensor temperature to 0 degree Celsius, which can further reduce temperature-related noise. So We believe that the HS can be regarded as ground truth with confidence.

Considering the feasibility of proposed set-up, the used object lens is FUJINON HF12.5HA-1S, with a focal length of 12.5mm, allowing 2/3 inch sensor. The active image resolution is 640x480, that is, 5.7mm $\times$ 4.3 mm, given the effective pitch size of 9um $\times$ 9um. Since the relay lens has a ratio of 1:1, the actual angle of view is 26.1 degrees (horizontal) and 19.5 degrees (vertical). At a distance of 10m, the field of view is 4.6m (horizontal) and 3.4m (vertical). So, we believe this setting is practical. Moreover, lens distortion can be an issue when coupled with RS effects, yet we tried to alleviate this by using a high-quality FUJINON lens.

A.3. Experimental Data

A.3.1 RD-VFI Dataset

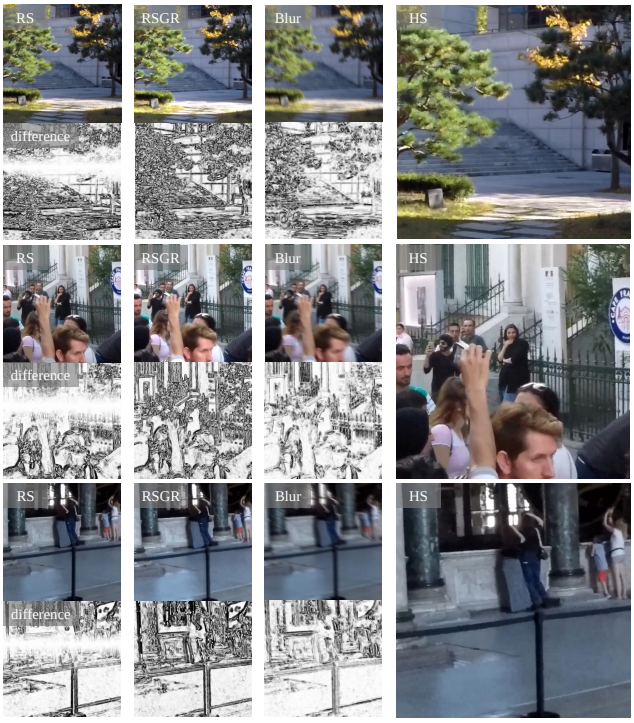
We have elaborated the collection process of our real data, RD-VFI in Section 3. By our quad-axis imaging system, we collected 133 video quadruples and each quadruple has 60×3 degraded frames and 600 sharp HS frames. The presented data samples of RD-VFI are in Fig. A.2. In difference map, the black area denotes larger differences while smaller one is in gray. Notably, the RS has sharper content with tilt effects. Blur and brightness of RSGR are highly related to image rows which is quite different with row-independent blur in GS images.

A.3.2 GOPRO-VFI Dataset

Although, we collected real-world data to contrast effects of different shutter mode induced degradations towards VFI. As complementary part, we also validated the findings on synthetic data, GOPRO-VFI. Following the convention [8, 10, 1], synthesis process is also grounded on GOPRO data [7] consisting of 33 videos with resolution of 1280×720. Each video clip contains about 1200 consecutive frames at 240fps. For benefits of generating more realistic effects, GOPRO is firstly interpolated at $\times64$ using an off-the-shelf video interpolation algorithm [5].

Fig. A.3 illustrates our synthesizing method. The RS videos are synthesized by sequentially copying a row of pix-

els from high-speed videos and blur generated by averaging them, as in previous works [6, 10]. The RSGR synthesizing process is similar to that of bur, but has two different parts: 1) Different rows of an RSGR frame are contributed by variant numbers of high-speed frames; 2) All used HS frames except for the first one are multiplied by a factor δ, which determines the ratio between readout time and the first scanline's exposure duration. The generation process of RS, RSGR and blur videos are strictly aligned to each frame ensuring they capture identical contents of the scene. In practice, we centrally crop frames to 512, and set $T = \tau = 512, \delta = 0.001$. Finally, we have 33 videos with three degradation counterparts and corresponding high-speed frames. Fig. A.4 exhibits some examples of GOPRO-VFI dataset.



Extended Figure A.4: **Data samples of different degradations under the same scene from our GOPRO-VFI**.

B. Additional Results

B.1. Experimental Results on GOPRO-VFI

As supplemental demonstration of conclusion drawn on real-data RD-VFI, we also conduct experiments on synthetic dataset GOPRO-VFI. The qualitative and quantitative results are presented in Fig. A.5 and Tab. A.1, Tab. A.2.

B.2. Video Reconstruction of Compared Methods

In Fig.6, we present the reconstructed consecutive frames of our PMBNet. Here, the corresponding counterparts of DeMFI [8], RSSR [2] and CVR [3] are shown in Fig. B.6, Fig. B.7 and Fig. B.8, respectively.

B.3. Additional Qualitative Results

We present additional qualitative comparisons on RD-VFI dataset in Fig. B.9 – Fig. B.13.

References

[1] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4641–4650, 2021. 1, 2

[2] Bin Fan and Yuchao Dai. Inverting a rolling shutter camera: bring rolling shutter images to high framerate global shutter video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4228–4237, 2021. 3, 4

[3] Bin Fan, Yuchao Dai, Zhiyuan Zhang, Qi Liu, and Mingyi He. Context-aware video reconstruction for rolling shutter cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17572–17582, 2022. 3, 4

[4] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[5] 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. 2

[6] Peidong Liu, Zhaopeng Cui, Viktor Larsson, and Marc Pollefeys. Deep shutter unrolling network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5941–5949, 2020. 3

[7] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 2

[8] Jihyong Oh and Munchurl Kim. Demfi: deep joint deblurring and multi-frame interpolation with flow-guided attentive correlation and recursive boosting. In *European Conference on Computer Vision*, pages 198–215. Springer, 2022. 2, 3, 4

[9] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European Conference on Computer Vision*, pages 184–201. Springer, 2020. 1

[10] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8174–8182, 2018. 2, 3

[11] Zhihang Zhong, Yinqiang Zheng, and Imari Sato. Towards rolling shutter correction and deblurring in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9219–9228, 2021. 1

Extended Table A.1: **Quantitative comparisons on RS and RSGR mode of GOPRO-VFI dataset**. We compare our model with state-of-the-art methods for degraded video frame interpolation. The performance is measured with mean PSNR, SSIM and LPIPS. The numbers in bold represent the best performance. To better compare all methods, we provide the evaluation metrics of correction, VFI (8 times interpolation) and average. $B0, B1$ denote initial performance of inputs. Correction metrics are computed from B_0, B_1 and S_0, S_1 while interpolation part is obtained by averaging all intermediate frames S_t.

Methods	RS Mode									RSGR Mode								
	Correction			VFI (×8)			Average			Correction			VFI (×8)			Average		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
B_0, B_1	22.95	0.8167	0.1649	-	-	-	-	-	-	18.05	0.7271	0.1941	-	-	-	-	-	-
RSSR [2]	23.41	0.8300	**0.0641**	23.05	0.8202	0.0893	23.13	0.8224	0.0837	19.24	0.7552	**0.0989**	19.13	0.7582	**0.1233**	19.15	0.7575	**0.1179**
CVR [3]	23.51	0.8339	0.0796	23.51	0.8340	**0.0803**	23.51	0.8340	**0.0801**	17.59	0.6612	0.1856	17.41	0.6504	0.1895	17.45	0.6528	0.1886
DeMFI [8]	23.46	0.8277	0.1348	23.42	0.8258	0.1535	23.43	0.8262	0.1494	22.11	0.7980	0.2310	22.46	0.8139	0.2003	22.38	0.8104	0.2071
PMBNet	**24.06**	**0.8471**	0.1158	**23.98**	**0.8450**	0.1199	**24.00**	**0.8455**	0.1190	**23.44**	**0.8051**	0.2583	**23.87**	**0.8203**	0.2357	**23.77**	**0.8169**	0.2407

Extended Table A.2: **Quantitative comparisons** on blur mode of GOPRO-VFI dataset.

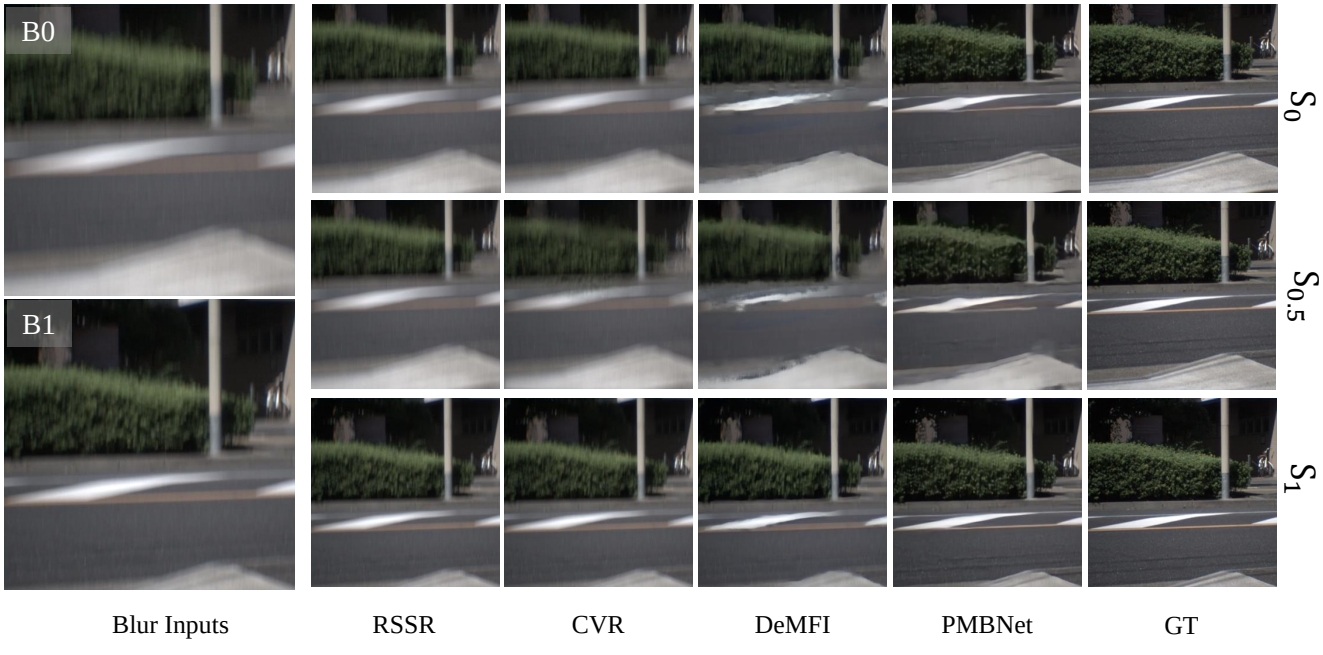
Method	Deblurring			VFI (x8)			Average		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
B_0, B_1	26.80	0.8810	0.2292	-	-	-	-	-	-
RSSR [2]	26.81	0.8805	0.2293	25.83	0.8643	0.2532	26.05	0.8679	0.2479
CVR [3]	26.51	0.8751	0.1787	26.32	0.8724	0.1786	26.36	0.8730	0.1786
DeMFI [8]	27.73	0.9007	0.1435	27.57	0.8993	0.1367	27.61	0.8996	0.1382
PMBNet	**31.52**	**0.9442**	**0.0817**	**31.26**	**0.9421**	**0.0795**	**31.32**	**0.9426**	**0.0800**

Extended Figure A.5: **Visual comparison on GOPRO-VFI.** We compare VFI results by different methods with RS, RSGR and GS blur degradations, respectively. In each mode, we present the results of $S_{2/8}$ (top row) and $S_{6/8}$ (bottom row).

Extended Figure B.6: **Reconstructed consecutive frames from two degraded inputs by using DeMFI.** We present the multiple intermediate frames at different time generated by three types of shutter induced degradations. They are temporally located at t = [0, 1) with stride of 0.1 and arranged in two rows from left to right. *Best viewed in zoom*.

Extended Figure B.7: **Reconstructed consecutive frames from two degraded inputs by using RSSR.** We present the multiple intermediate frames at different time generated by three types of shutter induced degradations. They are temporally located at t = [0, 1) with stride of 0.1 and arranged in two rows from left to right. *Best viewed in zoom*.

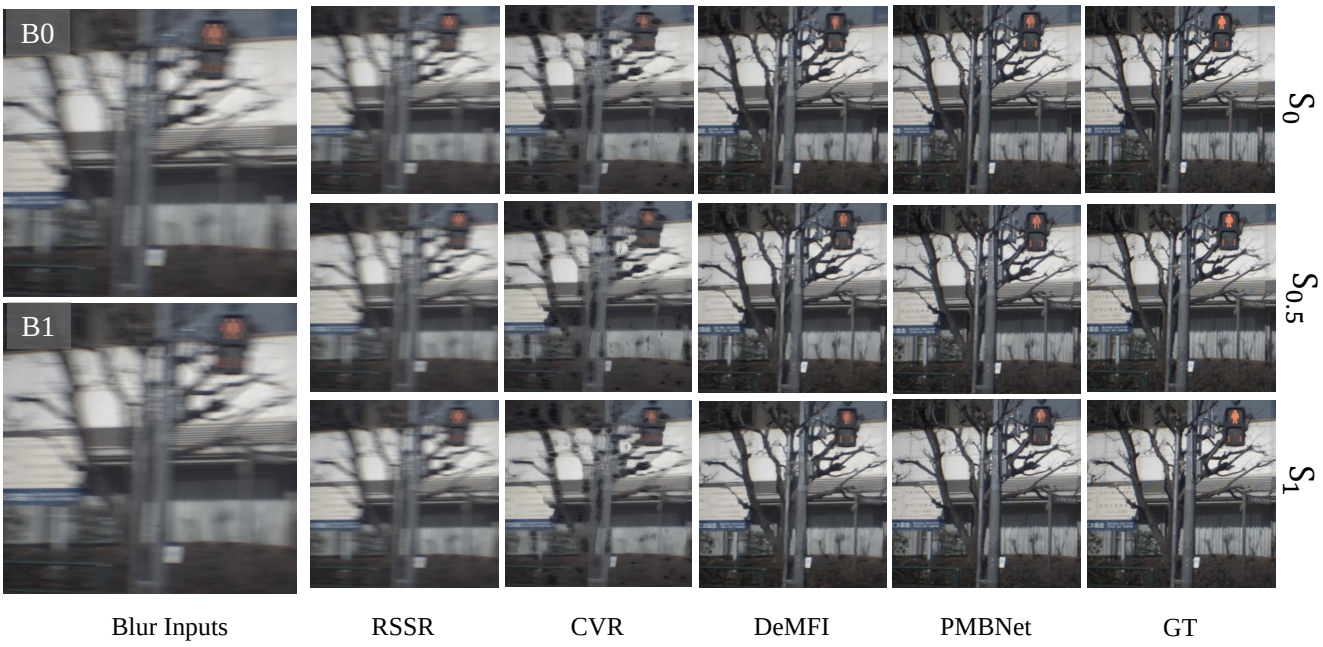
Extended Figure B.8: **Reconstructed consecutive frames from two degraded inputs by using CVR.** We present the multiple intermediate frames at different time generated by three types of shutter induced degradations. They are temporally located at t = [0, 1) with stride of 0.1 and arranged in two rows from left to right. *Best viewed in zoom*.



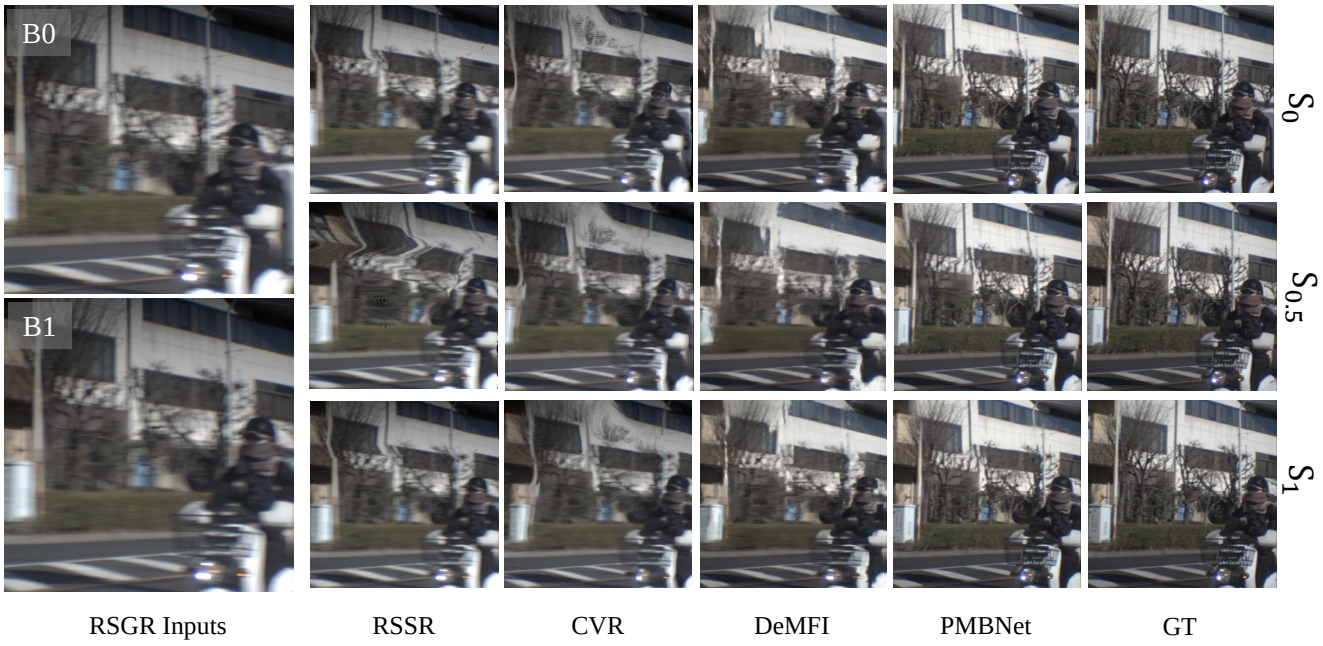
Extended Figure B.9: **Additional qualitative comparisons on blur mode of RD-VFI.**

Extended Figure B.10: **Additional qualitative comparisons on RS mode of RD-VFI.**

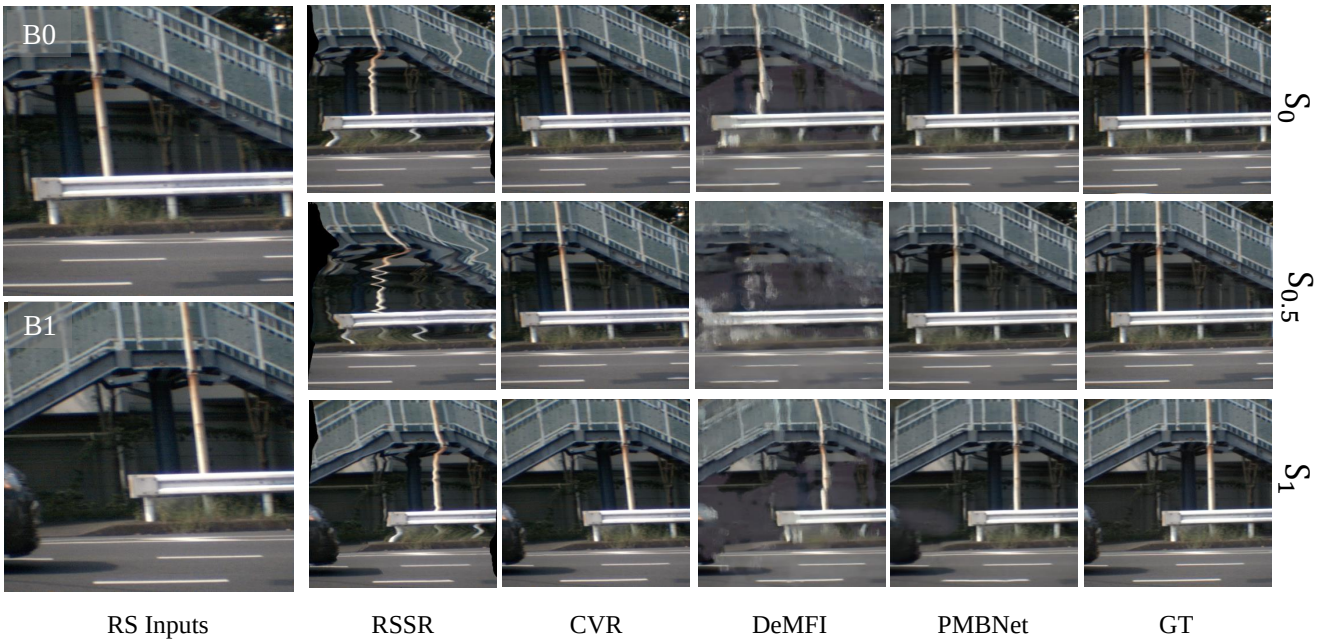
Extended Figure B.11: **Additional qualitative comparisons on RSGR mode of RD-VFI.**



Extended Figure B.12: **Additional qualitative comparisons on blur mode of RD-VFI.**



Extended Figure B.13: **Additional qualitative comparisons on RSGR mode of RD-VFI.**



Extended Figure B.14: **Additional qualitative comparisons on RS mode of RD-VFI.**