Context-Aware Video Reconstruction for Rolling Shutter Cameras –Supplementary Material–

Bin Fan Yuchao Dai* Zhiyuan Zhang Qi Liu Mingyi He
School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China
{binfan,zhangzhiyuan}@mail.nwpu.edu.cn, {daiyuchao,liuqi,myhe}@nwpu.edu.cn

Abstract

In this supplementary material, we first derive a general parametric form of bilateral motion field (BMF) by considering the readout time ratio, and then justify our problem setup. Next, we provide thorough analyses to our ABMF model, occlusion reasoning, and motion enhancement. Afterward, we show additional experimental results on RS correction, intermediate flow, generalization, and GS video recovery, which fully demonstrates the superiority of our pipeline. We also include a video demo to present dynamic GS video reconstruction results. Additional details of the loss function are then added. Furthermore, we report a partial ablation study of CVR. At last, several failure cases are given to look forward to possible future research.*

1. Instructions on Problem Setup

In this section, we show a detailed derivation of the general parameterization of BMF in the time dimension, followed by an explanation of our problem setup.

1.1. General formulation of BMF

We first give a brief description of the connection between the motion field $\mathbf{U}_{0\to s}$ and the optical flow $\mathbf{F}_{0\to 1}$ by accounting for the first RS frame $\mathbf{I}_0^r$ as an example. Since this does not contain our contribution, we only give the necessary details to follow the derivation below. More details of this connection can be found in [2]. Suppose that to estimate $\mathbf{U}_{0\to s}$ that warps each pixel (*e.g.* $\mathbf{x}$ in scanline κ) of $\mathbf{I}_0^r$ to the GS counterpart corresponding to its scanline s, this connection under the constant velocity motion model can be formulated as:

$$\mathbf{U}_{0\to s}(\mathbf{x}) = \mathbf{C}_{0\to s}(\mathbf{x}) \cdot \mathbf{F}_{0\to 1}(\mathbf{x}), \tag{1}$$

where

$$\mathbf{C}_{0\to s}(\mathbf{x}) = \frac{\gamma(s-\kappa)(h-\gamma\pi_v)}{h^2}, \tag{2}$$

denotes the forward correction map. Here, $\gamma \in (0, 1]$ is the readout time ratio [17], h is the number of scanlines, and π_v represents the latent inter-GS-frame vertical optical flow.

Next, we extend it to the time domain and derive a more general formulation than that in the main paper, *i.e.* γ will be taken into account. According to the definition in the main paper, the one-to-one correspondence between time t and scanline s will satisfy

$$t = \frac{\gamma}{h}\left(s - \frac{h}{2}\right). \tag{3}$$

It is easy to verify that the central scanlines of $\mathbf{I}_0^r$ and $\mathbf{I}_1^r$ correspond to time instances 0 and 1, respectively. Note that the first scanline of $\mathbf{I}_0^r$ will coincide with time $-\frac{\gamma}{2}$.

Assume that $t \in [0, 1]$ corresponds to the scanline s to be restored, and $\tau_0 \in [-\frac{\gamma}{2}, \frac{\gamma}{2}]$ corresponds to the exposure time of scanline κ of $\mathbf{I}_0^r$, we can obtain $t - \tau_0 \triangleq \frac{\gamma(s-\kappa)}{h}$. Further combining with Eq. (2) yields

$$\mathbf{C}_{0\to t}(\mathbf{x}) = \frac{(t-\tau_0)(h-\gamma\pi_v)}{h}. \tag{4}$$

Similarly, the backward correction map that accounts for the second RS frame $\mathbf{I}_1^r$ with $\tau_1 \in [1-\frac{\gamma}{2}, 1+\frac{\gamma}{2}]$ can be defined as:

$$\mathbf{C}_{1\to t}(\mathbf{x}) = \frac{(\tau_1-t)(h+\gamma\pi_v')}{h}. \tag{5}$$

Note that Eqs. (4) and (5) model the bilateral correction map through the time paradigm in a general sense. In this way, the general parametric form of BMF is modeled. This generality is reflected by the fact that, unlike Eq. (5) in the main paper, we get $\tau_0 \in [-\frac{\gamma}{2}, \frac{\gamma}{2}]$ and $\tau_1 \in [1-\frac{\gamma}{2}, 1+\frac{\gamma}{2}]$ instead of $\tau_0 \in [-0.5, 0.5]$ and $\tau_1 \in [0.5, 1.5]$. This is because we assume $\gamma = 1$ in our problem setup. In the following, we will explain the feasibility of this setup.

1.2. Feasibility analysis of our problem setup

Our problem setup with $\gamma = 1$ is based on three main reasons. Firstly, the Carla-RS and Fastec-RS datasets pro-

*Corresponding author

posed by [8] are the only available RS correction benchmark datasets. And they are constructed by using the assumption of $\gamma = 1$. Since we leverage these two datasets to train our network, this setup can also be considered as we use $\gamma = 1$ in advance. Secondly, as manifested by [2, 3, 8] and our experiments, the trained deep learning model based on these two datasets can be successfully generalized to RS data acquired by real cameras (*i.e*. γ may not be equal to 1). Also, γ is directly set to 1 in [4, 18] to correct real RS images, which can avoid the non-trivial readout calibration. These studies demonstrate that assuming $\gamma = 1$ is generally feasible for modeling RS correction problems. Thirdly, this setup can facilitate temporally tractable frame interpolation between two consecutive RS images. Note that we derive the general parameterization of BMF in Subsec. 1.1, which will help the future exploration of more general RS-based video reconstruction tasks (*e.g*. with extremely small γ).

2. Additional Architecture Analyses

In this section, we provide more in-depth analyses of our method in terms of the proposed ABMF model, time-aware occlusion reasoning and motion enhancement architectures.

2.1. Further analysis on ABMF model

Here, we show the rationality of our ABMF model proposed in Subsection 3.1 of the main paper, *i.e*., we need to verify that $h \pm |\pi_v| \approx h$ (equivalently, $\frac{|\pi_v|}{h} \approx 0$). To this end, we define $\frac{|\pi_v|}{h}$ as the vertical pixel displacement ratio. From [2], one can get $\frac{|\pi_v|}{h} = |\mathbf{f}_v/(h + \mathbf{f}_v)|$, where $\mathbf{f}_v$ denotes the inter-RS-frame vertical optical flow and h is the number of image rows. We employ the state-of-the-art optical flow estimation pipeline RAFT [14] to obtain the optical flow map between two adjacent RS frames. Then, we calculate the average value and standard deviation of $\frac{|\pi_v|}{h}$ in each image of the test sets of Carla-RS and Fastec-RS datasets. We plot their respective statistics in Fig. 1. One can observe that the latent inter-GS-frame vertical optical flow value is usually much smaller than the number of image rows, *i.e*. the proposed ABMF model is concise and reasonable. Note that $\frac{\gamma|\pi_v|}{h} \leq \frac{|\pi_v|}{h}$, which indicates that our ABMF model is also valid under the general formulation of Subsec. 1.1. Furthermore, as illustrated in Fig. 2, the ABMF-based RSSR* tends to have misaligned errors and unsmooth artifacts at local boundaries (*e.g*. depth variation, slight blurring, monotonous texture, *etc*.) due to the isotropic approximation of ABMF. Fortunately, the experimental results demonstrate that our *CVR** (*i.e*. combining ABMF with the GS frame refinement module) enhances local details and improves image quality in a coarse-to-fine manner, which can serve as an effective and efficient baseline for RS-based video reconstruction. Note that our *CVR* can further improve the fidelity and authenticity of the recovered GS images because a better initial estimate is provided by using the network-based bilateral motion field (NBMF).

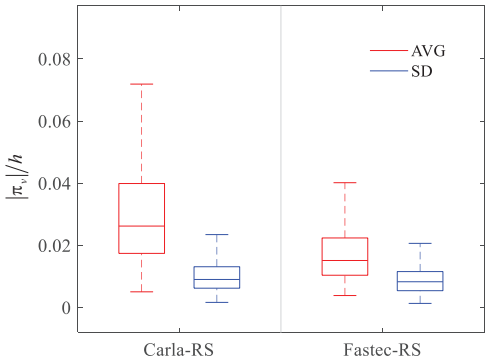


Figure 1. Statistical results of the vertical pixel displacement ratio $\frac{|\pi_v|}{h}$ under the Carla-RS and Fastec-RS datasets. Red is the average value (AVG) and blue is the standard deviation (SD).

2.2. Further analysis on occlusion reasoning

As can be seen from Fig. 3, the severe occlusion exists in the pool at the lower-left corner of the RS images (*cf*. blue circle). Since RSSR [2] only uses the contents of a single RS image to synthesize the corresponding GS image, a mass of occluded black holes inevitably appear (*cf*. red circles). In contrast, we mitigate this struggle by effectively aggregating contextual information through occlusion inference. Interestingly, in the examples of time $t = 0.5$ shown in Fig. 3, the estimated bilateral occlusion masks can vividly reflect the human intuitive observation discussed in [3], *i.e*. the first and second rolling shutter images contribute greatly to the lower and upper parts of the latent GS image at time $t = 0.5$, respectively. Meanwhile, the GS image corresponding to time $t = 0$ or $t = 1$ will be more convinced by the RS image that is closer in time, which reasonably follows the RS imaging mechanism. In a nutshell, our method can restore high-quality GS images with richer details, enhancing the visual experience. Note that our method can also model temporal abstractions in an end-to-end manner, which allows adaptively generating time-aware occlusion masks to obtain GS images at arbitrary times.

2.3. Further analysis on motion enhancement

We further investigate the effectiveness of our motion enhancement layer in Fig. 4, taking the correction to time $t = 1$ as an example. As illustrated by the red boxes, the motion enhancement scheme facilitates the quality of the bilateral motion field. As a result, the local image details (*e.g*. object-specific motion boundaries, small errors, *etc*.) are refined so as to encourage subsequent contextual aggregation. Combined with the proposed contextual consistency constraint, it can promote high-fidelity GS frame synthesis with the assistance of bilateral occlusion masks.

Figure 2. Example results of the effectiveness of our ABMF model. Since the ABMF model ignores the depth variations, the ABMF-based RSSR* may encounter misaligned errors and unsmooth artifacts at motion boundaries, while the (NBMF-based) RSSR [2] can alleviate these problems to some extent, but still not as well as it could be. Combined with the GS frame refinement module, ABMF provides concise and tractable benefits for GS video recovery, while NBMF can yield better initialization to generate higher fidelity GS images.

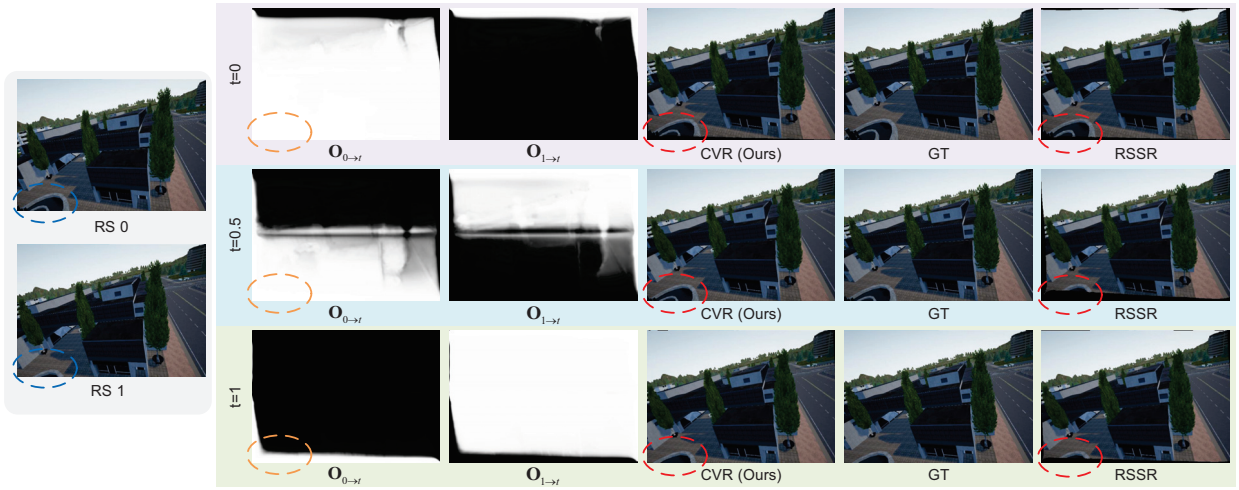


Figure 3. Example results of the effectiveness of our occlusion reasoning. We show GS frame recovery at times 0, 0.5 and 1, respectively. The brighter the color in the bilateral occlusion mask, the higher the credibility. Our method can adaptively and efficiently reason about complex occlusions and temporal abstractions, leading to visually more satisfactory GS reconstruction results than RSSR [2].

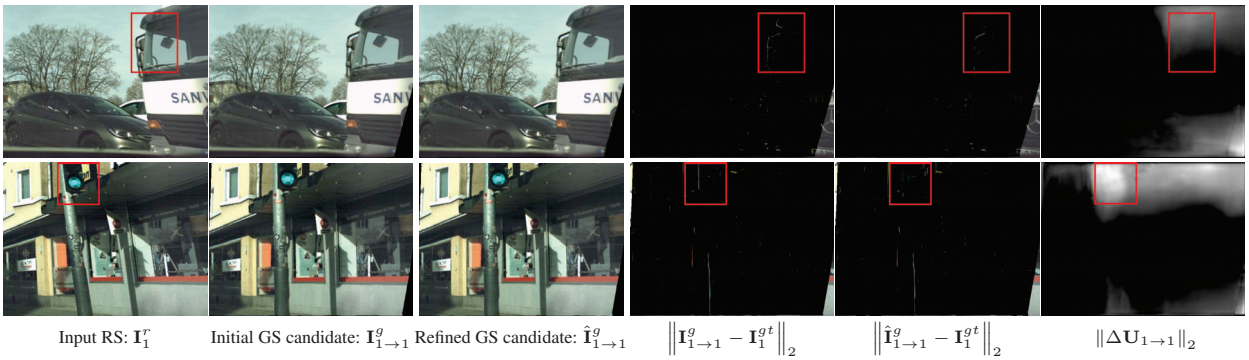
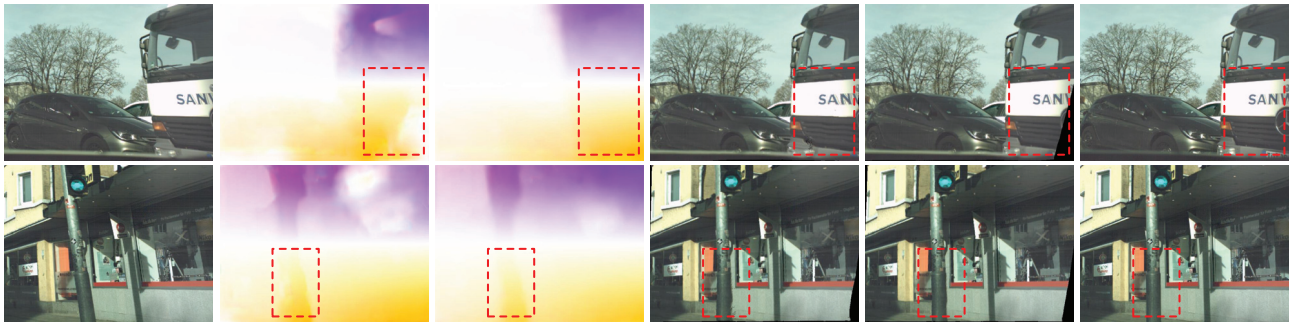


Figure 4. Example results of the effectiveness of our motion enhancement. The second to fifth columns show the initial intermediate GS frame candidates, the refined intermediate GS frame candidates, and their absolute differences with corresponding ground-truth, respectively. The sixth column indicates the mean of the BMF residual map (the brighter a pixel, the bigger the motion enhancement). Our *CVR* effectively enhances ambiguous motion boundaries for more accurate contextual alignment.



Input RS 1 | Intermediate flow (RSSR) | Intermediate flow (CVR) | Estimated GS (RSSR) | Estimated GS (CVR) | Ground-truth

Figure 5. Visual results of the intermediate flow estimation. The estimated GS images in the fourth and fifth columns are obtained by warping the input RS frames according to the intermediate flows in the second and third columns, respectively. Our *CVR* estimates the intermediate flow with clearer motion boundaries than RSSR [2] and thus generates more accurate and sharper GS content.

Table 1. Quantitative comparisons on recovering GS images at time step $t = 1$. The numbers in **red** and blue represent the best and second-best performance. In addition to the SOTA quantification performance for GS image recovery at time $t = 0.5$, our method also obtains almost consistent best metrics at time $t = 1$. Note that not only these, high-quality GS video frames corresponding to any time $t \in [0, 1]$ can be accurately estimated by our method.

Method	PSNR↑ (dB)			SSIM↑		LPIPS↓	
	CRM	CR	FR	CR	FR	CR	FR
DeepUnrollNet [8]	27.86	27.54	**27.02**	0.829	0.828	0.0555	**0.0791**
RSCD [16]	-	-	24.84	-	0.778	-	0.1070
RSSR [2]	29.36	26.57	24.89	0.900	0.824	0.0553	0.1109
*CVR** (Ours)	28.28	28.19	26.58	0.912	0.833	0.0444	0.1014
CVR (Ours)	**29.41**	**29.19**	26.67	**0.915**	**0.838**	**0.0403**	0.1011

*: *applying our proposed approximated bilateral motion field (ABMF) model.*

3. Additional Experimental Results

In this section, we present more qualitative and quantitative experimental results on effect removal, intermediate flow, and generalization, *etc*. Furthermore, a video demo is included to show the dynamic results of reconstructing slow-motion GS video from two consecutive RS frames.

3.1. RS effect removal

First, we visualize the intermediate flow in Fig. 5 and compare it with the SOTA RS-based video reconstruction method RSSR [2]. In contrast to RSSR, our pipeline generates intermediate flows with clearer motion boundaries for more accurate frame interpolation due to motion interpretation and occlusion reasoning. Then, we report more RS effect removal results in Fig. 6 and Fig. 8 by comparing with the off-the-shelf video frame interpolation (VFI) and RS correction algorithms. Finally, in Table 1, we give quantitative comparison results of GS image recovery at time $t = 1$. In addition to the superior RS effect removal performance at time $t = 0.5$, our pipeline also significantly surpasses RSSR at time $t = 1$. In summary, these experimental results consistently demonstrate that the proposed method has superior RS effect removal capabilities, successfully restoring higher fidelity global shutter video frames with fewer artifacts and richer details.

3.2. Generalization on other real data

To evaluate the generalization performance of the proposed method on real rolling shutter images, we utilize the data provided by [17] and [5], in which the hand-held cameras move quickly in the real world to capture real RS image sequences. As shown in Fig. 9, our *CVR* and *CVR** can effectively and robustly remove the RS effect to obtain consistent distortion-free images, which validates the excellent generalization performance of our method in practice.

3.3. GS video reconstruction demo

We attach a supplementary video demo_video.mp4 to dynamically demonstrate the GS video reconstruction results. In the video, we show the $10\times$ temporal upsampling results, *i.e*. evenly interpolating 11 intermediate GS frames corresponding to time steps 0, 0.1, 0.2, ..., 0.9, 1. In essence, our method is capable of generating GS videos with arbitrary frame rates. Note that except for times 0, 0.5, and 1, our method has not been fed with GS images of other time instances during training. More qualitative results on RS correction datasets [8] and real RS data [5, 17] can be seen in the supplementary video. With these examples, we can conclude that our method not only achieves the state-of-the-art RS effect removal performance that is significantly better than competing methods, but also has the superior ability to recover high-quality and high-framerate GS videos.

Figure 6. Visual examples against off-the-shelf VFI approaches (*i.e.* BMBC [12], DAIN [1], and Cascaded method). Although the cascaded method can compensate the drawback of the VFI method that cannot remove RS artifacts, it is also prone to local errors due to error accumulation, as shown by the red circles.

4. Details of Loss Function

Assuming that T GS images at time instances $\{t_i\}_1^T, t_i \in [0,1]$ are to be recovered to supervise the training of our model and that $\mathbf{I}_{t_i}^{gt}$ is the corresponding ground-truth (GT) GS image, our loss function $\mathcal{L}$ is a linear combination of the reconstruction loss $\mathcal{L}_r$, perceptual loss $\mathcal{L}_p$, contextual consistency loss $\mathcal{L}_c$, and total variation loss $\mathcal{L}_{tv}$, *i.e.*

$$\mathcal{L} = \lambda_r \mathcal{L}_r + \mathcal{L}_p + \lambda_c \mathcal{L}_c + \lambda_{tv} \mathcal{L}_{tv}, \tag{6}$$

where λ_r, λ_c and λ_{tv} are hyper-parameters. The pixel intensities of images are normalized.

The *reconstruction loss* $\mathcal{L}_r$ models the pixel-wise $\mathcal{L}_1$ loss betweeen the final GS frame prediction and the corresponding ground-truth, given by

$$\mathcal{L}_r = \frac{1}{T} \sum_{i=1}^{T} \left\| \hat{\mathbf{I}}_{t_i}^{g} - \mathbf{I}_{t_i}^{gt} \right\|_1 . \tag{7}$$

The *perceptual loss* $\mathcal{L}_p$ contributes to produce fine details and improves the perceptual quality of the final intermediate GS frame [7] by

$$\mathcal{L}_p = \frac{1}{T} \sum_{i=1}^{T} \left\| \phi\left(\hat{\mathbf{I}}_{t_i}^{g}\right) - \phi\left(\mathbf{I}_{t_i}^{gt}\right) \right\|_1 , \tag{8}$$

where ϕ is the conv4_3 features of the pre-trained VGG16 network [13], as widely used in [6, 8, 10].

The *contextual consistency loss* $\mathcal{L}_c$ encourages the alignment of the refined intermediate GS frame candidates and their ground-truth frame at time t_i. This can also facilitate the final enhanced BMF to reason about the underlying occlusions and the object-specific motion boundaries, which are crucial for the final GS frame synthesis. Specifically, we define $\mathcal{L}_c$ as:

$$\mathcal{L}_c = \frac{1}{2T} \sum_{i=1}^{T} \left(\left\| \hat{\mathbf{I}}_{0\to t_i}^{g} - \mathbf{I}_{t_i}^{gt} \right\|_1 + \left\| \hat{\mathbf{I}}_{1\to t_i}^{g} - \mathbf{I}_{t_i}^{gt} \right\|_1 \right). \tag{9}$$

The *total variation loss* $\mathcal{L}_{tv}$ enforces piecewise smoothness in the final enhanced BMF [3, 9], *i.e.*

$$\mathcal{L}_{tv} = \frac{1}{2T} \sum_{i=1}^{T} \left(\left\| \nabla \hat{\mathbf{U}}_{0\to t_i} \right\|_2 + \left\| \nabla \hat{\mathbf{U}}_{1\to t_i} \right\|_2 \right). \tag{10}$$

Table 2. Ablation results for *CVR** architecture on $\mathcal{M}_A$ and $\mathcal{G}$.

Settings	PSNR↑ (dB)			SSIM↑	
	CRM	CR	FR	CR	FR
RAFT-based	30.40	29.91	27.67	0.914	0.835
Freeze $\mathcal{M}_A$	31.69	31.53	28.51	**0.927**	0.843
$\mathbf{T} \cdot \Delta\mathbf{U}$	31.15	30.95	28.01	0.916	0.831
w/o $\Delta\mathbf{U}$	31.61	31.41	27.99	0.925	0.831
w/o $\mathbf{O}$	30.96	30.80	23.89	0.913	0.804
full model	**31.82**	**31.60**	**28.62**	**0.927**	**0.845**

5. Ablations of the Proposed CVR*

Additionally, we report the impact of different network architecture designs on our *CVR** in Table 2 by referring to the GS images at time $t = 0.5$. Ablation results that are almost consistent with *CVR* in the main paper can be obtained, which fully demonstrates the validity of the network architecture we used.

6. Failure Cases

We have discussed that our pipeline may have blending and ghosting artifacts in image regions such as

Figure 7. Failure cases in some challenging image areas. The white pillars and car tails lack texture and are thus prone to aliasing artifacts.

low/weak/repetitive textures. We reckon this is because our method exploits image-based warping, and thus potential gross errors of the estimated BMF in these challenging regions can easily lead to contextual misalignment. In fact, this is a common challenge for the current RS correction method based on image warping, *e.g.* [2, 17–19]. We show visual results of the failure cases in Fig. 7. Similar to the training process of VFI methods [6, 11, 15], it will likely be helpful to use more GT GS images at different time steps to supervise the training of our network. In the future, we also plan to improve the BMF estimation or design feature-based aggregation schemes to ameliorate this weakness.

References

[1] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019. 5

[2] Bin Fan and Yuchao Dai. Inverting a rolling shutter camera: bring rolling shutter images to high framerate global shutter video. In *Proceedings of IEEE International Conference on Computer Vision*, pages 4228–4237, 2021. 1, 2, 3, 4, 6, 7

[3] Bin Fan, Yuchao Dai, and Mingyi He. Sunet: symmetric undistortion network for rolling shutter correction. In *Proceedings of IEEE International Conference on Computer Vision*, pages 4541–4550, 2021. 2, 5, 7

[4] Bin Fan, Ke Wang, Yuchao Dai, and Mingyi He. Rolling-shutter-stereo-aware motion estimation and image correction. *Computer Vision and Image Understanding*, 213:103296, 2021. 2

[5] Per-Erik Forssén and Erik Ringaby. Rectifying rolling shutter video from hand-held devices. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 507–514, 2010. 4, 8

[6] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: high quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018. 5, 6

[7] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of European Conference on Computer Vision*, pages 694–711, 2016. 5

[8] Peidong Liu, Zhaopeng Cui, Viktor Larsson, and Marc Pollefeys. Deep shutter unrolling network. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5941–5949, 2020. 2, 4, 5

[9] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of IEEE International Conference on Computer Vision*, pages 4463–4471, 2017. 5

[10] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1701–1710, 2018. 5

[11] Avinash Paliwal and Nima Khademi Kalantari. Deep slow motion video reconstruction with hybrid imaging system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1557–1569, 2020. 6

[12] Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. Bmbc: bilateral motion estimation with bilateral cost volume for video interpolation. In *Proceedings of European Conference on Computer Vision*, pages 109–125, 2020. 5

[13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015. 5

[14] Zachary Teed and Jia Deng. Raft: recurrent all-pairs field transforms for optical flow. In *Proceedings of European Conference on Computer Vision*, pages 402–419, 2020. 2

[15] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In *Proceedings of Advances in Neural Information Processing Systems*, volume 32, 2019. 6

[16] Zhihang Zhong, Yinqiang Zheng, and Imari Sato. Towards rolling shutter correction and deblurring in dynamic scenes. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9219–9228, 2021. 4

[17] Bingbing Zhuang, Loong-Fah Cheong, and Gim Hee Lee. Rolling-shutter-aware differential sfm and image rectification. In *Proceedings of IEEE International Conference on Computer Vision*, pages 948–956, 2017. 1, 4, 6, 7, 8

[18] Bingbing Zhuang and Quoc-Huy Tran. Image stitching and rectification for hand-held cameras. In *Proceedings of European Conference on Computer Vision*, pages 243–260, 2020. 2, 6, 7

[19] Bingbing Zhuang, Quoc-Huy Tran, Pan Ji, Loong-Fah Cheong, and Manmohan Chandraker. Learning structure-and-motion-aware rolling shutter correction. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4551–4560, 2019. 6

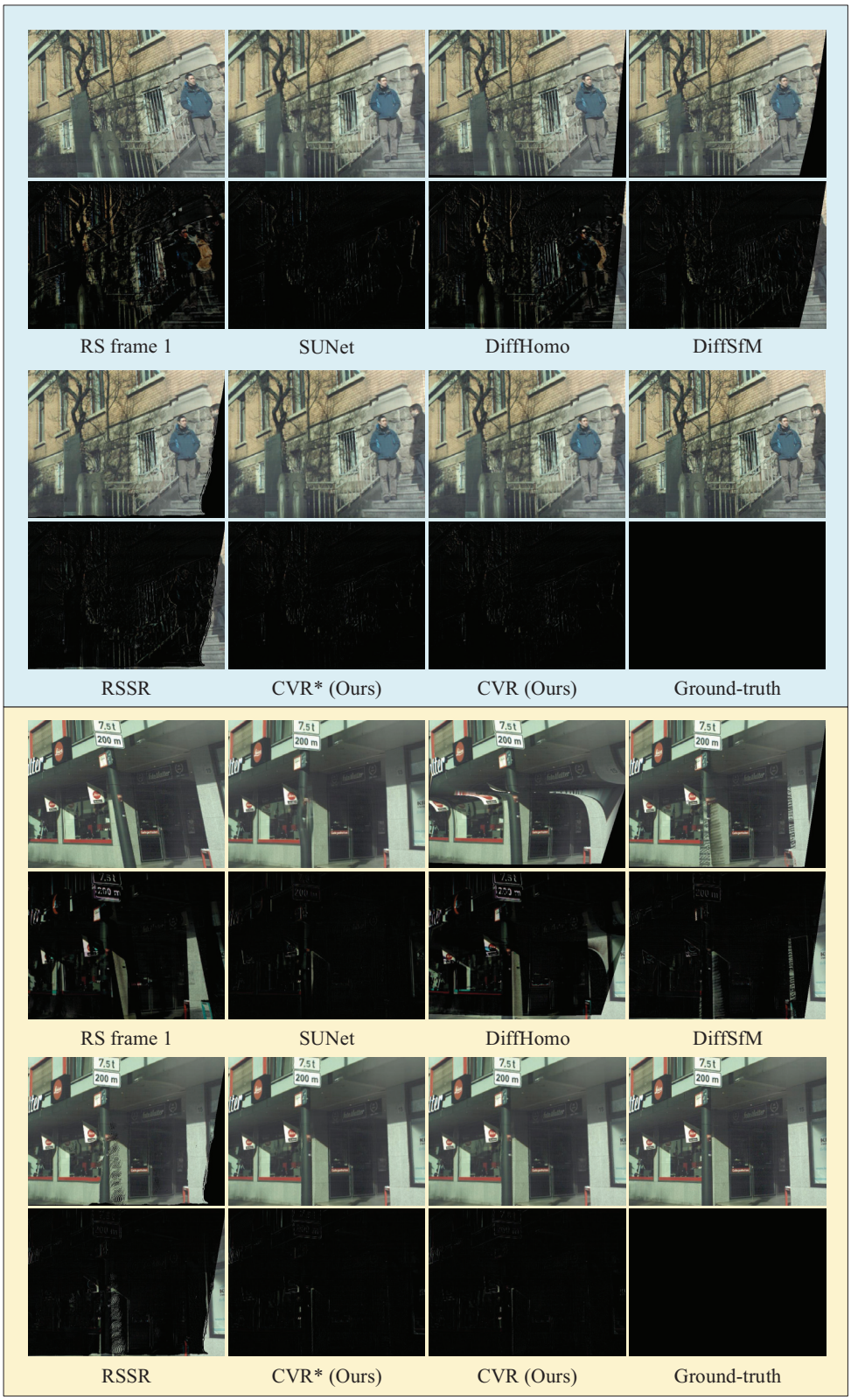


Figure 8. Rolling shutter effect removal examples against competing approaches (*i.e.* SUNet [3], DiffHomo [18], DiffSfM [17], and RSSR [2]). Even columns: Absolute difference between the corrected global shutter image and the corresponding ground-truth.

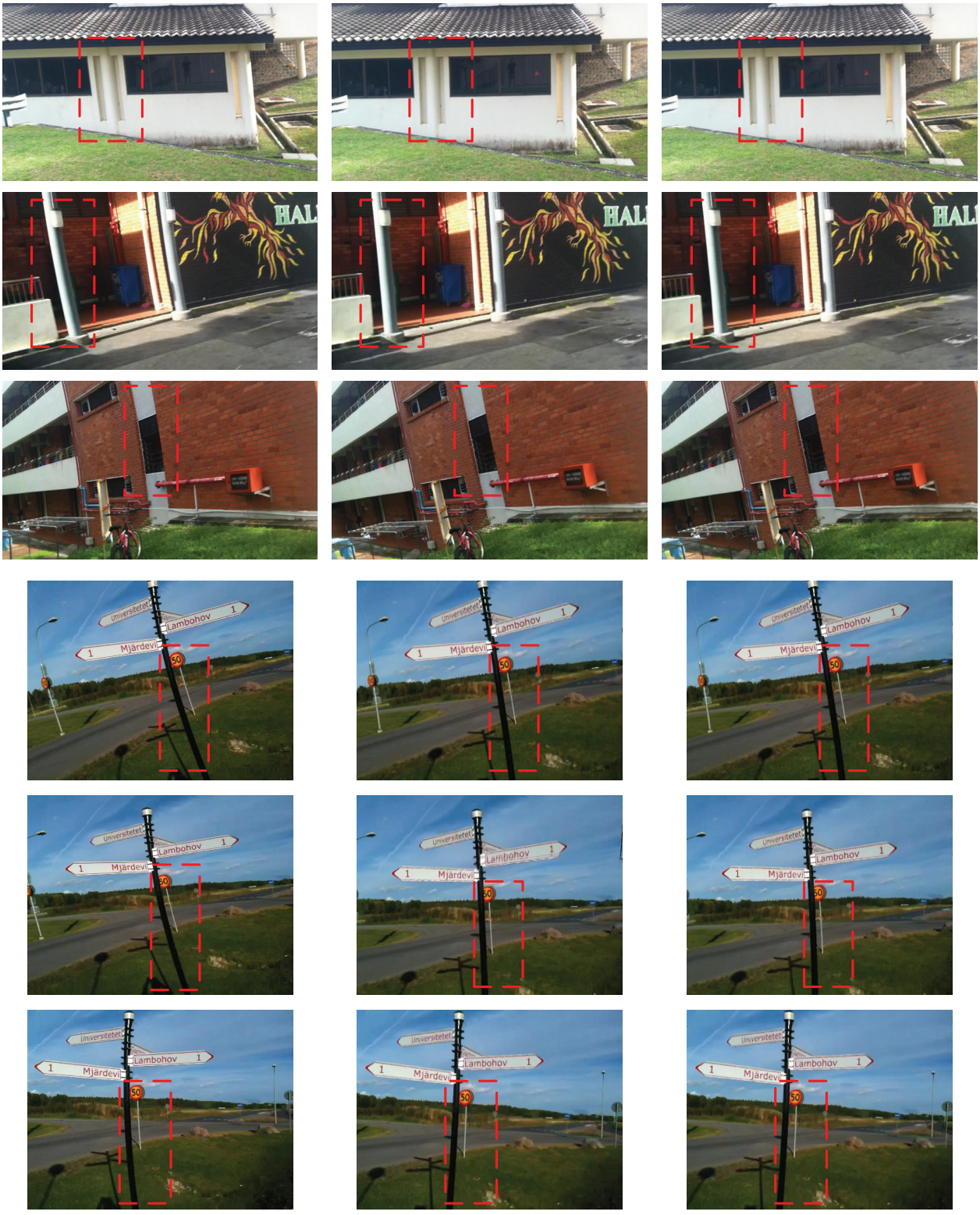


Figure 9. Generalization results on real rolling shutter data with noticeable rolling shutter artifacts. The data in the first three rows are from [17], and the last three rows are from [5]. Consistent and high-quality correction results are obtained by our *CVR* and *CVR**.