Inverting a Rolling Shutter Camera: Bring Rolling Shutter Images to High Framerate Global Shutter Video (Supplementary Material)

Bin Fan Yuchao Dai*
School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China
binfan@mail.nwpu.edu.cn, daiyuchao@nwpu.edu.cn

Abstract

In this supplementary material, we show the superiority of our method in RS effect removal in Fig. 1, and further provide detailed ablation studies on network $\mathcal{F}$, network $\mathcal{U}$, and loss function $\mathcal{L}$. Afterward, we demonstrate the advantages of our method over the state-of-the-art video frame interpolation methods. And more experimental analyses on DeepUnrollNet [4] are carried out. We also show the generalization results of our method by using other real RS images. Furthermore, more qualitative results and a video demo are included to prove the effectiveness of our method in recovering high framerate GS video frames. In the end, we derive the RS-aware backward warping model that accounts for the second RS frame, and then summarize the details of our loss function.

1. Ablation Studies

We present a series of ablation studies of our architecture design in terms of backbone networks $\mathcal{F}$ and $\mathcal{U}$ and loss function $\mathcal{L}$. We will explain each of them in detail below.

1.1. Ablation on the selection and training strategy of network $\mathcal{F}$

We first replace PWC-Net [10] with the state-of-the-art optical flow estimation baseline RAFT [11]. Then, we analyze the influence of different training strategies of network $\mathcal{F}$, including parameter freezing and model initialization during training, *i.e.*,

- $\mathcal{F}$-*Scra*: We initialize $\mathcal{F}$ from scratch and optimize it with the whole model.
- $\mathcal{F}$-*Pret*: We initialize $\mathcal{F}$ from the pre-trained model provided by [10] and optimize it with the whole model.

The quantitative results are summarized in Table 1 and Table 2. We can observe that RAFT contributes slightly worse than PWC-Net to the high framerate GS video extraction in our implementation. Freezing the network parameters when using RAFT can significantly improve performance, but it has the opposite benefit when combined with PWC-Net. When initializing from the pre-trained model, especially with the addition of fine-tuning, our method shows a substantial performance improvement. After jointly optimizing the whole network together with

Inputs (Overlay) DeepUnrollNet [4] DiffHomo [14] DiffSfM [13] RSSR (Ours) Ground truth

Figure 1: A difficult example for joint RS correction and temporal super-resolution (towards the first scanline of the second frame). Our method fully explores the underlying RS geometry and generates a set of high-quality GS results, in spite of the road sign that is subject to large RS artifacts.

Table 1: Ablation study on the selection and training strategy of network $\mathcal{F}$. We employ different optical flow estimation baselines (RAFT [11] and PWC-Net [10]), while testing the effect of freezing their parameters during training.

RAFT	PWC-Net	Freeze	PSNR↑		SSIM↑	
			CRM	FR	CR	FR
✓		✓	29.81	20.63	**0.87**	0.77
✓		×	26.75	20.38	0.84	0.71
	✓	✓	29.44	20.64	0.86	0.77
	✓	×	**30.17**	**21.26**	**0.87**	**0.78**

Table 2: Effectiveness of different components of our model on the Carla-RS dataset.

	PSNR↑		SSIM↑	LPIPS↓
	CRM	CR	CR	CR
$\mathcal{F}$-Scra	27.37	24.22	0.80	0.0804
$\mathcal{F}$-Pret	29.89	24.61	0.86	0.0697
w/o $\mathbf{T}$	25.43	22.55	0.82	0.1116
w/o $\Delta\mathbf{F}$	29.12	24.29	0.85	0.0725
w/o $\mathcal{L}_r$	29.44	24.35	0.86	0.0713
w/o $\mathcal{L}_p$	29.82	24.61	0.86	0.0706
w/o $\mathcal{L}_s$	29.28	24.44	0.86	0.0725
full model	**30.17**	**24.78**	**0.87**	**0.0695**

PWC-Net, the overall performance is further improved, which is better capable of exploiting the concealed motion between scanlines as well as the scene structure.

1.2. Ablation on the design of network $\mathcal{U}$

We further investigate the contribution of each component in network $\mathcal{U}$ as follows:

- w/o $\mathbf{T}$: We remove the normalized scanline offset in Eq. (19) of the main manuscript, and replace the *Sigmoid* function with the *Tanh* function in network $\mathcal{U}$ to uniformly map the correlation factor prediction of each pixel to the interval of $(-1, 1)$.
- w/o $\Delta\mathbf{F}$: We remove the optical flow residual estimation layer in network $\mathcal{U}$, *i.e.*, $\Delta\mathbf{F}_{1\to2} = \Delta\mathbf{F}_{2\to1} = \mathbf{0}$ in Eq. (20) of the main manuscript.

We report the results in Table 2 and Fig. 2. One can see that the explicit constraint of the normalized scanline offset benefits the learning the scanline-dependent nature of the RS undistortion flow, which is consistent with the observation in [4]. Also, adding the optical flow residual estimation layer is effective to facilitate the edge alignment and improve the robustness of the proposed model in the extreme case, thereby recovering more complete image details.

1.3. Ablation on the loss function

We show the results of training our models under different loss function settings in Table 2. We remove each loss term from the overall loss function $\mathcal{L}$ respectively. Without

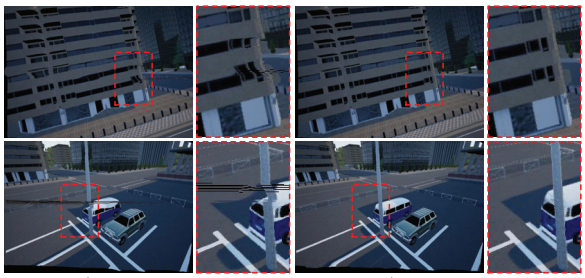


Figure 2: The optical flow residual estimation layer can effectively alleviate the artifacts and holes at the boundaries caused by optical flow misalignment, resulting in higher quality results.

$\mathcal{L}_w$ indicates freezing the parameters of PWC-Net in Table 1. Forcing the smoothness of the estimated flows has a particularly positive effect on improving the performance. Our loss function is effective as the performance of adopting all loss terms is the best.

2. Versus Video Frame Interpolation Methods

The current video frame interpolation algorithms, *e.g.*, BMBC [6] and DAIN [1], have a common implicit assumption that the camera uses a global shutter, where the pixel displacement is controllable and located in the corresponding optical flow. Specifically, linearly scaling the optical flow between 0 and 1 to approximate the required intermediate pixel displacement in order to warp input images. In contrast, to correct the RS image, as shown in Eq. (13) in the main manuscript, the pixel displacement is neither a linear function of scanline time (including complex RS geometry) nor within the corresponding optical flow (*i.e.*, the length of the RS undistortion flow may be larger than that of the optical flow, or its direction may be opposite to the optical flow), involving inherent non-local operations. Therefore, because of inherent flaws in the network architectures, the existing video frame interpolation algorithms are incapable of eliminating the RS effect effectively. We validate this argument in Fig. 3, which also highlights the superiority of our RSSR method in recovering high-quality distortion-free GS video frames.

Furthermore, we conduct experiments to compare with the two-stage approach, *i.e.*, given three consecutive RS images, we first obtain two corrected GS images in sequence using DeepUnrollNet [4], and then interpolate the GS image corresponding to the first scanline of the third RS image using DAIN [1]. As visualized in Fig. 4, this two-stage approach suffers from serious blur artifacts as well as local inaccuracies. Moreover, it is computationally inefficient. For instance, to recover 960 GS video frames, the two-stage approach takes about 5 minutes (at least 0.3 seconds to generate a frame with DAIN), while our RSSR method takes

Figure 3: Visual results against video frame interpolation algorithms (BMBC [6] and DAIN [1]) to generate an intermediate frame corresponding to the intermediate time of two consecutive RS frames. Only our proposed RSSR method can successfully remove RS artifacts.

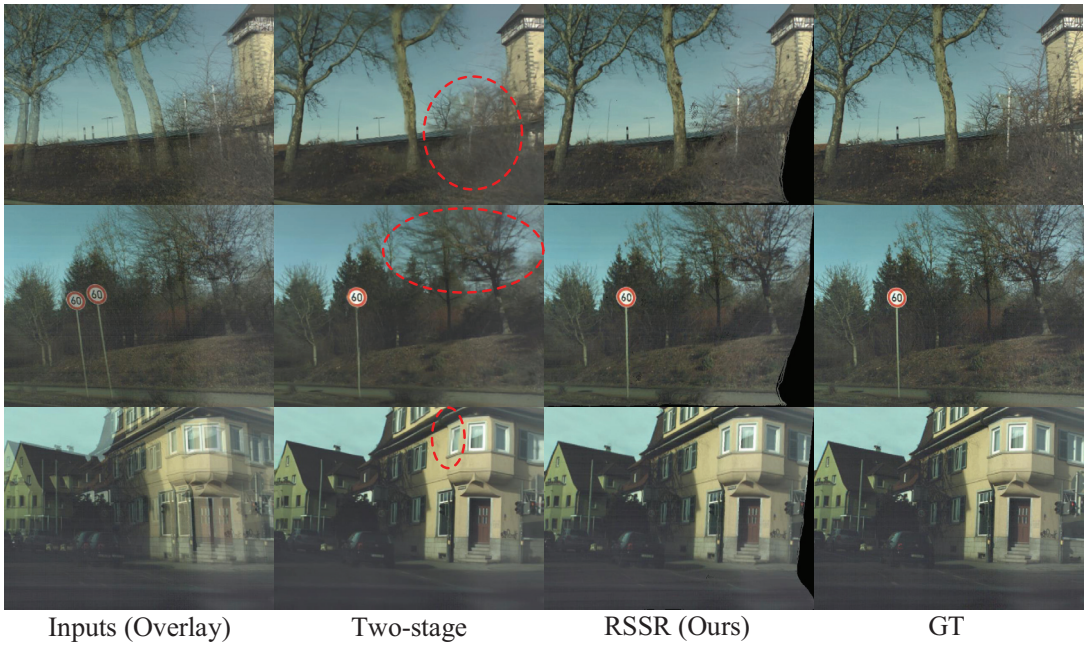


Figure 4: Visual results against the two-stage approach: perform RS correction first, then perform video frame interpolation.

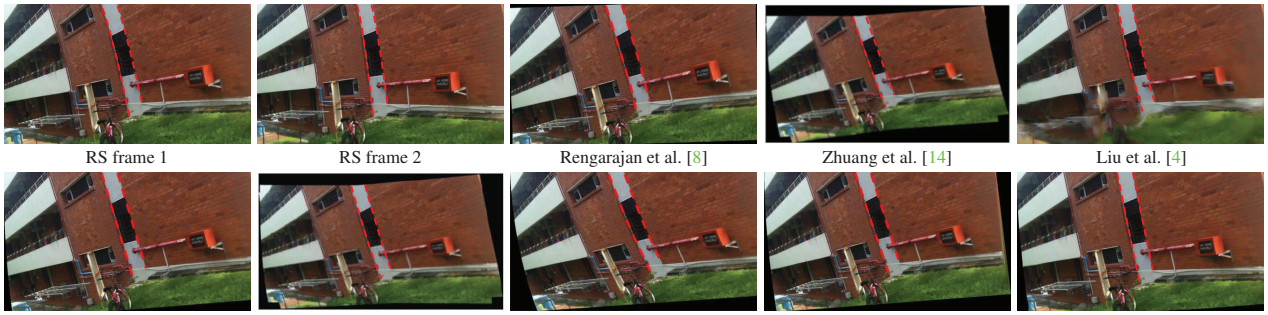


Figure 5: Qualitative comparison of image correction results on real data with obvious RS distortion provided by [13]. Our pipeline can effectively estimate plausible GS images.

Table 3: Quantitative comparisons of the performance between our approach and DeepUnrollNet [4] in recovering GS images corresponding to the middle scanline of the second RS frame. Note that, in other chapters and the main manuscript, all competing methods refer to the first scanline of the second RS frame.

Method	PSNR↑			SSIM↑		LPIPS↓	
	CRM	CR	FR	CR	FR	CR	FR
DeepUnrollNet [4]	27.86	**27.54**	**27.02**	0.829	0.828	0.0555	**0.0791**
RSSR (Ours)	**29.36**	26.57	25.01	**0.900**	**0.834**	**0.0553**	0.0817

a total of 1.8 seconds. Since our method solves the RS correction and temporal super-resolution simultaneously, it achieves excellent performance in terms of both accuracy and efficiency.

3. Instruction on Comparison With DeepUnrollNet [4]

In our evaluation, in order to be consistent with [13–15] (*i.e.*, recovering the GS image corresponding to the first scanline of the second frame), all competing results relate to the first scanline of the second frame. Note that we retrain DeepUnrollNet [4] to adapt to this task for fair comparison. However, the original publication of [4] is mainly used to restore a GS image corresponding to the middle scanline of the second frame, so we additionally add quantitative and qualitative results under the middle scanline of the second frame, as shown in Table 3 and Fig. 6. One can see that, except for some occluded edges, our method is comparable or superior to DeepUnrollNet in returning the GS image corresponding to the middle scanline of the second frame. Note also that our method can restore the GS image corresponding to any scanline *without demanding access to the supervision of the corresponding GT GS images*. In addition, our method is more satisfactory when restoring the GS image corresponding to the first scanline of the second frame. However, DeepUnrollNet is limited to generate a single reliable GS frame that corresponds to the middle scanline of the second frame, and has poor adaptability to the recovery of GS images of other scanlines, even if the corresponding precious GT GS supervision is provided.

4. Generalization to other Real RS Data

Our learning-based model is trained on the Carla-RS dataset, within which the RS artifacts are mainly caused by uniform camera motion. We apply our method to real RS images provided by [13] and show example results in Fig. 5. The results reveal that our proposed method owns good generalization ability and can recover visually compelling GS images, due to learning the underlying RS geometry.

5. More Qualitative Experimental Results

We provide more qualitative results on the Carla-RS and Fastec-RS datasets, as shown in Figs. 7 and 8. Also, an enlarged result is reported in Fig. 1. Compared with the off-the-shelf RS correction algorithms, including SMARSC [15], DeepUnrollNet [4], DiffHomo [14], and DiffSfM [13], our pipeline effectively restores higher quality GS images. Particularly, our RSSR model can also recover smooth and high framerate GS video sequences.

6. RSSR Video Demo

We attach a video named "RSSR_Video.mp4" to show the dynamic results of our method in simultaneous RS correction and temporal super-resolution. In principle, we can produce videos with arbitrary frame rates. Our RSSR network learns to solve the complex RS geometry embedded in the consecutive RS frames, so it can robustly and accurately recover photorealistic time-continuous GS images. Overall, our method not only has the advantage of RS correction at a specific scanline time, but also has the superior ability to restore GS images at any scanline time.

Figure 6: Visual comparison of GS images corrected to the middle scanline of the second RS frame. At this time, the corresponding plausible GS image can be reconstructed by DeepUnrollNet [4] and our method. Since we have not developed a modulated image decoder as in [4], our method cannot fill the occluded regions. Our RSSR method, however, is able to recover GS video images at any scanline, which is far beyond the reach of DeepUnrollNet.

7. RS-Aware Backward Warping Model

To formulate the RS-aware backward warping accounting for frame 2, the backward inter-frame camera velocities $(\mathbf{v}', \boldsymbol{\omega}')$ should obey: $\mathbf{v}' = -\mathbf{v}$ and $\boldsymbol{\omega}' = -\boldsymbol{\omega}$. Let Z' denote the depth of each pixel $\mathbf{x}'$ in frame 2 and $(\mathbf{f}'_u, \mathbf{f}'_v)$ the backward optical flow from frame 2 to frame 1. In complete analogy with the RS forward motion parameterizations in Section 3 of the main manuscript, we again derive the relative motion between scanline s_1 of frame 1 and scanline s_2 of frame 2 as:

$$\begin{aligned} \mathbf{v}_{s_2 s_1} &= (\lambda_2^{s_2} - \lambda_1^{s_1})\, \mathbf{v}', \\ \boldsymbol{\omega}_{s_2 s_1} &= (\lambda_2^{s_2} - \lambda_1^{s_1})\, \boldsymbol{\omega}', \end{aligned} \tag{1}$$

where

$$\lambda_1^{s_1} = \tfrac{\gamma s_1}{h}, \quad \lambda_2^{s_2} = 1 + \tfrac{\gamma s_2}{h}. \tag{2}$$

Note that $\mathbf{f}'_v = s_1 - s_2$. In the same way, we can obtain RS-aware backward warping model for the backward optical flow at pixel $\mathbf{x}'$ as:

$$\begin{bmatrix} \mathbf{f}'_u \\ \mathbf{f}'_v \end{bmatrix} = \alpha' \begin{bmatrix} \pi_u(\mathbf{v}', \boldsymbol{\omega}', \mathbf{x}', Z', f) \\ \pi_v(\mathbf{v}', \boldsymbol{\omega}', \mathbf{x}', Z', f) \end{bmatrix}, \tag{3}$$

where

$$\alpha' = 1 - \frac{\gamma \mathbf{f}'_v}{h} \tag{4}$$

represents the RS-aware backward interpolation factor under the constant velocity model.

Furthermore, we derive the RS-aware backward warping displacement vector, which transforms each RS pixel $\mathbf{x}'$ on κ-th scanline of frame 2 to arrive at a distortion-free frame

defined by the pose of s-th scanline of frame 2, as follows:

$$\begin{bmatrix} \mathbf{u}'_u \\ \mathbf{u}'_v \end{bmatrix} = \beta' \begin{bmatrix} \pi_u(\mathbf{v}', \boldsymbol{\omega}', \mathbf{x}', Z', f) \\ \pi_v(\mathbf{v}', \boldsymbol{\omega}', \mathbf{x}', Z', f) \end{bmatrix}, \tag{5}$$

where

$$\beta' = -\frac{\gamma(s-\kappa)}{h} \tag{6}$$

represents the RS-aware backward undistortion factor.

In a nutshell, Eq. (3) constrains the backward optical flow from frame 2 to frame 1 and Eq. (5) describes the backward RS undistortion flow from frame 2 to its scanline s. Just need to negative the readout time ratio γ, we can simply model the RS-aware backward warping. Other conclusions resemble those in forward warping. Note that the mutual conversion scheme between varying scanline-dependent RS undistortion flows is consistent in the forward and backward warpings, because they are independent of γ.

8. Details of the Loss Function

Given a pair of consecutive RS images $\mathbf{I}_r^1$ and $\mathbf{I}_r^2$, our network predicts the bidirectional optical flows $\mathbf{F}_{1\to2}$ and $\mathbf{F}_{2\to1}$, the bidirectional RS undistortion flows $\mathbf{U}_{1\to m}$ and $\mathbf{U}_{2\to m}$, and the target middle-scanline GS images $\mathbf{I}_g^1$ and $\mathbf{I}_g^2$. Let $\mathbf{I}_{gt}^1$ and $\mathbf{I}_{gt}^2$ denote the corresponding ground truth GS images. Our loss function $\mathcal{L}$ is a linear combination of the reconstruction loss $\mathcal{L}_r$, perceptual loss $\mathcal{L}_p$ [2], warping loss $\mathcal{L}_w$, and smoothness loss $\mathcal{L}_s$:

$$\mathcal{L} = \mu_r\mathcal{L}_r + \mu_p\mathcal{L}_p + \mu_w\mathcal{L}_w + \mu_s\mathcal{L}_s, \tag{7}$$

where μ_r, μ_p, μ_w and μ_s are hyper-parameters to balance different losses.

Reconstruction loss $\mathcal{L}_r$: We measure the pixel-wise reconstruction qualities of the corrected middle-scanline GS images as:

$$\mathcal{L}_r = \sum_{i=1}^{2} \left\| \mathbf{I}_g^i - \mathbf{I}_{gt}^i \right\|_1. \tag{8}$$

Perceptual loss $\mathcal{L}_p$: To mitigate the blur in the corrected middle-scanline GS images, we therefore use a perceptual loss $\mathcal{L}_p$ [2] to preserve details of the predictions and make estimated GS images sharper. Similar to [4], $\mathcal{L}_p$ is defined as:

$$\mathcal{L}_p = \sum_{i=1}^{2} \left\| \phi\left(\mathbf{I}_g^i\right) - \phi\left(\mathbf{I}_{gt}^i\right) \right\|_1, \tag{9}$$

where ϕ represents the $conv3\_3$ feature extractor of the VGG19 model [9].

Warping loss $\mathcal{L}_w$: Besides supervising the middle-scanline GS image predictions, we also introduce the warping loss $\mathcal{L}_w$ to maintain the qualities of the final bidirectional optical flows, defined as:

$$\mathcal{L}_w = \left\| \mathbf{I}_r^1 - g(\mathbf{I}_r^2, \hat{\mathbf{F}}_{1\to2}) \right\|_1 + \left\| \mathbf{I}_r^2 - g(\mathbf{I}_r^1, \hat{\mathbf{F}}_{2\to1}) \right\|_1, \tag{10}$$

where $\hat{\mathbf{F}}_{1\to2} = \mathbf{F}_{1\to2} + \Delta\mathbf{F}_{1\to2}$, $\hat{\mathbf{F}}_{2\to1} = \mathbf{F}_{2\to1} + \Delta\mathbf{F}_{2\to1}$, and $g(\cdot,\cdot)$ is the backward warping function.

Smoothness loss $\mathcal{L}_s$: At last, a smoothness term [5] is employed to enforce the smoothness of the bidirectional optical flows and bidirectional RS undistortion flows as:

$$\mathcal{L}_s = \sum_{i=1}^{2} \sum_{j=1, j\neq i}^{2} \left\| \nabla\hat{\mathbf{F}}_{i\to j} \right\|_2 + \left\| \nabla\mathbf{U}_{i\to m} \right\|_2. \tag{11}$$

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 the Conference on Computer Vision and Pattern Recognition*, pages 3703–3712. IEEE, 2019. 2, 3

[2] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 694–711. Springer, 2016. 6

[3] Yizhen Lao and Omar Ait-Aider. Rolling shutter homography and its applications. *Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2780–2793, 2021. 4

[4] Peidong Liu, Zhaopeng Cui, Viktor Larsson, and Marc Pollefeys. Deep shutter unrolling network. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 5941–5949. IEEE, 2020. 1, 2, 4, 5, 6

[5] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the International Conference on Computer Vision*, pages 4463–4471. IEEE, 2017. 6

[6] Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation. In *Proceedings of the European Conference on Computer Vision*, pages 109–125. Springer, 2020. 2, 3

[7] Vijay Rengarajan, Yogesh Balaji, and AN Rajagopalan. Unrolling the shutter: cnn to correct motion distortions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 2291–2299. IEEE, 2017. 4

[8] Vijay Rengarajan, Ambasamudram N Rajagopalan, and Rangarajan Aravind. From bows to arrows: rolling shutter rectification of urban scenes. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 2773–2781. IEEE, 2016. 4

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015. 6

[10] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 8934–8943. IEEE, 2018. 1, 2

[11] Zachary Teed and Jia Deng. Raft: recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*, pages 402–419. Springer, 2020. 1, 2

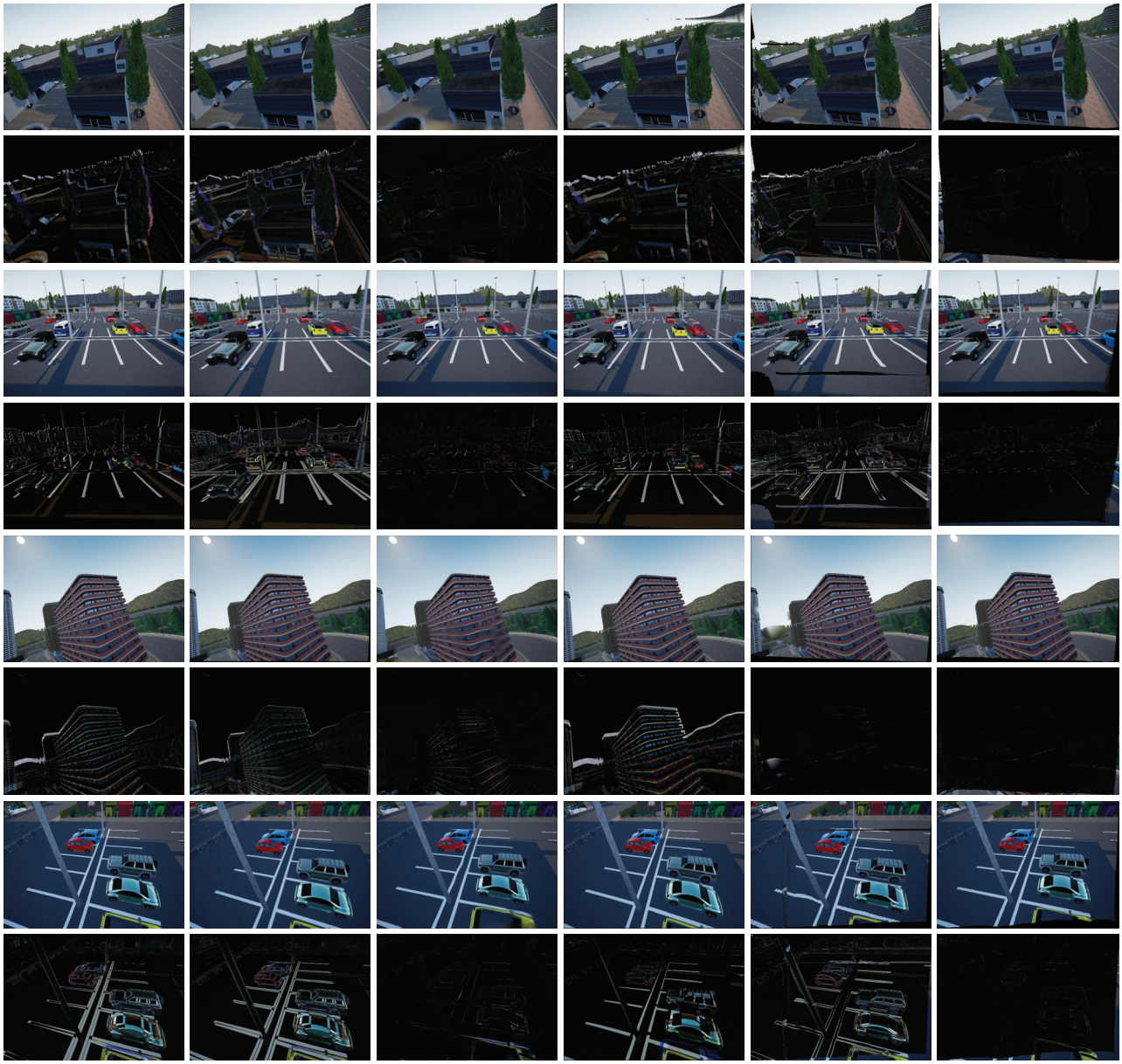


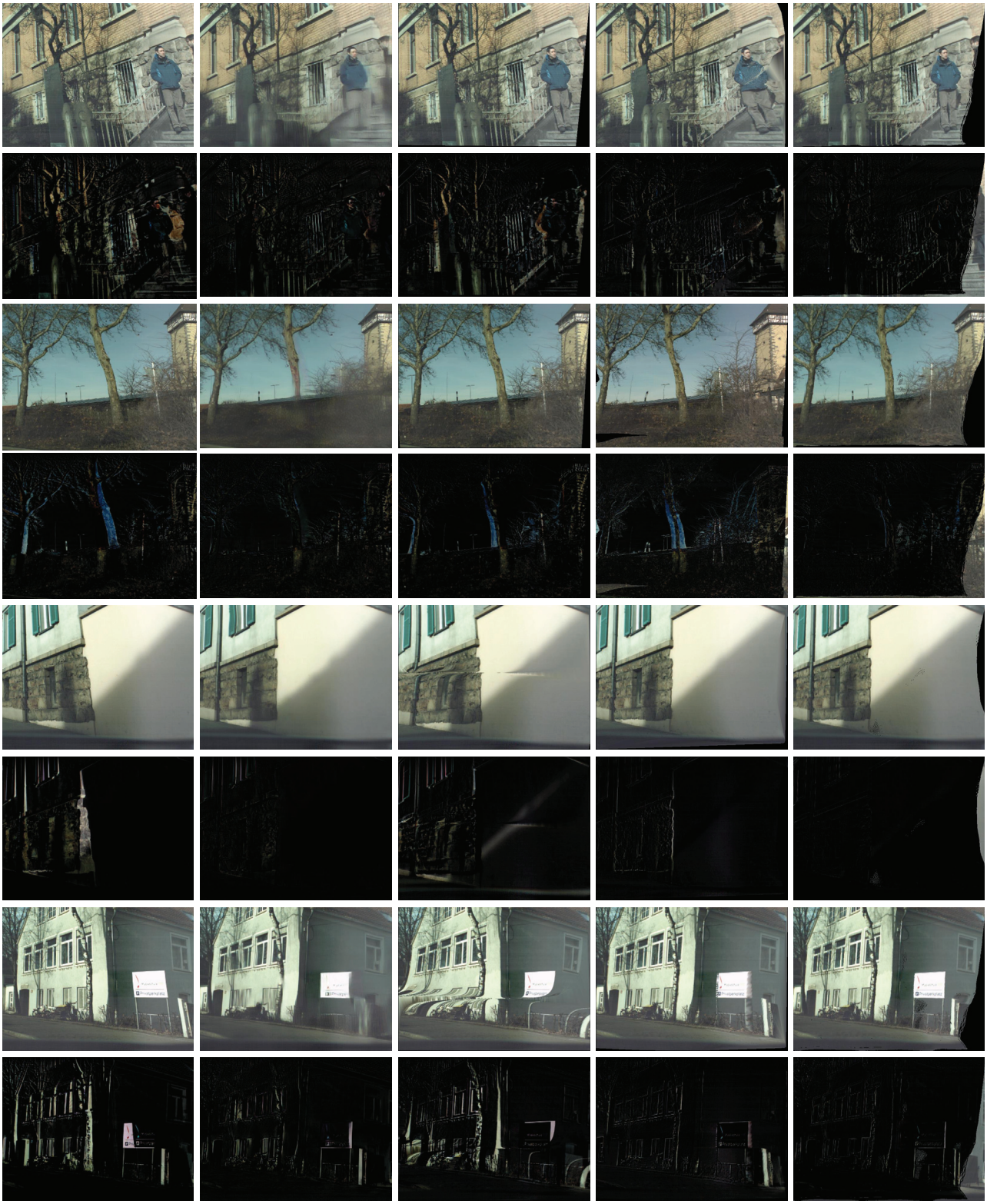
Figure 7: Qualitative results against baseline methods on the Carla-RS dataset. Even rows: Absolute difference between the corrected GS image and the corresponding GT GS image.

[12] Subeesh Vasu, Mahesh MR Mohan, and AN Rajagopalan. Occlusion-aware rolling shutter rectification of 3d scenes. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 636–645. IEEE, 2018. 4

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

[14] Bingbing Zhuang and Quoc-Huy Tran. Image stitching and rectification for hand-held cameras. In *Proceedings of the European Conference on Computer Vision*, pages 243–260. Springer, 2020. 1, 4

[15] Bingbing Zhuang, Quoc-Huy Tran, Pan Ji, Loong-Fah Cheong, and Manmohan Chandraker. Learning structure-and-motion-aware rolling shutter correction. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 4551–4560. IEEE, 2019. 4



RS frame 2 | DeepUnrollNet | DiffHomo | DiffSfM | RSSR (Ours)

Figure 8: Qualitative results against baseline methods on the Fastec-RS dataset. Even rows: Absolute difference between the corrected GS image and the corresponding GT GS image.