Joint Appearance and Motion Learning for Efficient Rolling Shutter Correction
–Supplementary Material–

Bin Fan⋆ Yuxin Mao⋆ Yuchao Dai† Zhexiong Wan Qi Liu
School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China
{binfan, maoyuxin, wanzhexiong}@mail.nwpu.edu.cn, {daiyuchao, liuqi}@nwpu.edu.cn

Abstract

In this supplementary material, we first introduce the network details of the joint appearance and motion decoder. Afterward, we provide more experimental results based on the real BS-RSC dataset [3]. We also include a video demo to present the rolling shutter correction results on consecutive sequences. Finally, we provide additional visual analysis of the intermediate processes and ablation experiments, followed by a discussion of the limitations of our method.

1. Network Details

To facilitate the overall understanding of the proposed joint learning mechanism, we do not show too many intermediate variables in Fig. 3 and Eq. 2 in the main text. Next, to better understand the intermediate computational process of our joint appearance and motion decoder, without loss of generality, we take the fourth level as an example for illustration, as shown in Fig. S1. Specifically, we describe in detail the computational process of equations

$$\begin{aligned}&[\tilde{U}^4_{g\to0}, \tilde{U}^4_{g\to1}, \tilde{G}^4, \tilde{h}^4] = \mathcal{D}^4([\hat{F}^4_0, \hat{F}^4_1, U^4_{g\to0}, U^4_{g\to1}, G^4, h^4]),\\ &[U^3_{g\to0}, U^3_{g\to1}, G^3, h^3] = Up([\tilde{U}^4_{g\to0}, \tilde{U}^4_{g\to1}, \tilde{G}^4, \tilde{h}^4]).\end{aligned} \tag{1}$$

Note that the tilde is used to represent the direct prediction of the current level, which is then upsampled for computation of the next level. For example, $\tilde{G}^5$ denotes the synthesized GS candidate at level 5, which is bilinearly upsampled to G^4 such that the joint learning is performed at level 4. And for network training, we supervise the bilateral undistortion fields $U^4_{g\to0}$, $U^4_{g\to1}$, the warping-based GS candidates $\hat{F}^4_0$, $\hat{F}^4_1$, and the synthesized GS candidate $\tilde{G}^4$ in Eq. (1). More specific details can be found in our code at `https://github.com/GitCVfb/JAMNet`.

2. Additional Experimental Results

In this section, we show more quantitative experimental results on GS image recovery, intermediate processes, and ablation experiments.

2.1. Visual results of GS image recovery

We report more real-world rolling shutter correction results in Fig. S3 and Fig. S4 by comparing with the off-the-shelf rolling shutter correction (RSC) algorithms, including AdaRSC [3], DeepUnrollNet [10], JCD [15], and SUNet [4]. As marked by the box, we can see that our JAMNet consistently outperforms the competing methods in visual appearance, successfully recovering higher-fidelity global shutter images with fewer artifacts and richer details. This is mainly attributed to the design of our single-stage architecture with joint appearance and motion learning. Moreover, we attach a supplementary video demo_video.mp4 to demonstrate the RSC results on consecutive RS video sequences (*e.g.*, Scenes 51, 54, 65 of the BS-RSC test set [3], involving noticeable RS artifacts).

2.2. Visualization of intermediate processes

As depicted in Fig. S5, we provide visual results of the intermediate process of our proposed JAMNet, including intermediate flows and warping-based GS candidates. As shown by the blue circles, there are different degrees of occlusion between the RS images and the target GS image, so simply warping one RS image leads to a significant lack of image content (see red circles). Our single-stage framework learns image appearance and pixel motion information simultaneously, and thus can adaptively and efficiently reason about complex occlusions to generate visually more satisfying RS correction results. Although we do not impose labeled auxiliary supervision (*e.g.* distillation loss [8, 11]) on the intermediate flows, thanks to our proposed joint learning mechanism, our baseline is capable of estimating a plausible bilateral undistortion field that better maintains the scanline-dependent property. In summary, by learning in a collaborative manner, the complementary

---

⋆ Equal contribution.
† Corresponding author.

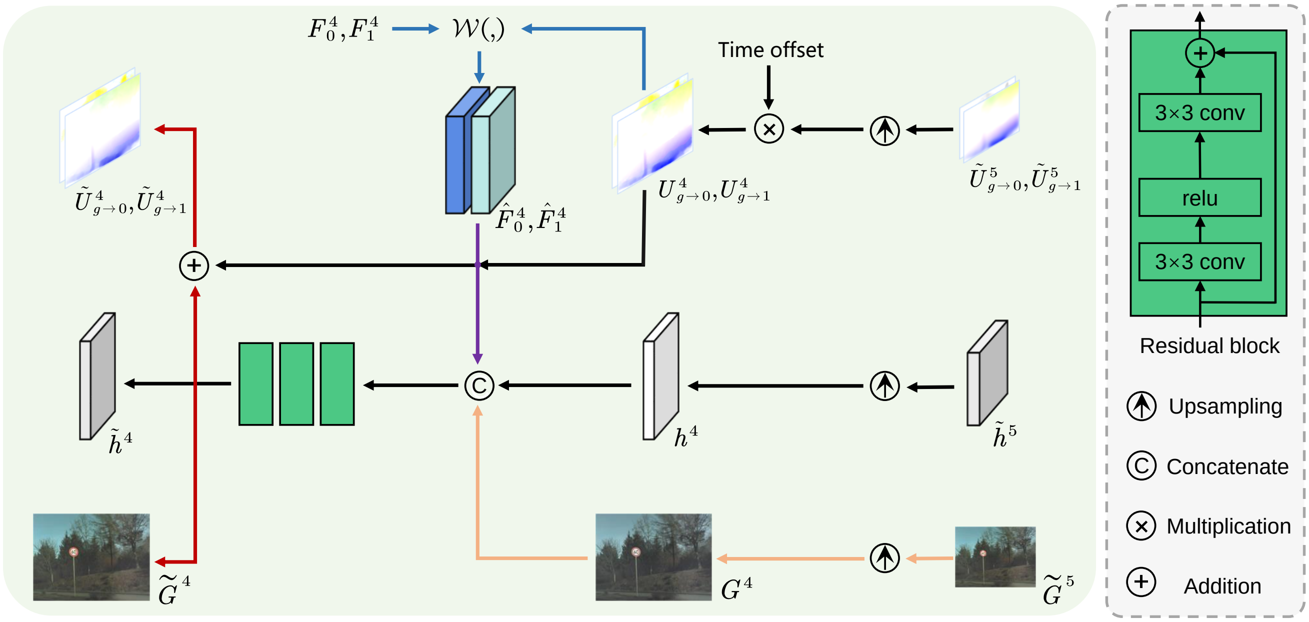


Figure S1. **Decoder details at level 4.** After upsampling the predictions $\tilde{U}^5_{g\to0}$, $\tilde{U}^5_{g\to1}$, $\tilde{h}^5$, and $\tilde{G}^5$ from the previous level, we obtain $U^4_{g\to0}$, $U^4_{g\to1}$, h^4, and G^4. Subsequently, in the warping branch, $U^4_{g\to0}$ and $U^4_{g\to1}$ are used to warp the feature map to obtain two warping-based GS candidates $\hat{F}_0^4$, $\hat{F}_1^4$. Finally, $U^4_{g\to0}$, $U^4_{g\to1}$, $\hat{F}_0^4$, $\hat{F}_1^4$, h^4, and G^4 are cascaded and fed into three residual blocks to output the predictions $\tilde{U}^4_{g\to0}$, $\tilde{U}^4_{g\to1}$, $\tilde{h}^4$, and $\tilde{G}^4$ for the current level in a joint learning manner. Note that inspired by [9, 13], the flow residual connections from $U^4_{g\to0}$, $U^4_{g\to1}$ are used to recover $\tilde{U}^4_{g\to0}$, $\tilde{U}^4_{g\to1}$. The whole process is optimized in an L-layer pyramid from coarse to fine, progressively obtaining the final GS image $\tilde{G}^1$ with the same resolution as the input.

motion estimation and appearance recovery in the RSC task can benefit each other. On top of combining the advantages of occlusion inference and context aggregation, our method finally recovers a photo-realistic GS image effectively.

2.3. Visualization of ablation experiments

In Table 3 of the main text, we show the quantitative results of the ablation experiments. To better understand the utility of each component, we visualize some representative ablation results in Fig. S6. "No synthesis" and "No warping" indicate that the synthesis branch and the warping branch are removed in the decoder, respectively. "No hidden" and "No $\mathcal{L}_{\text{mc}}$" represent the removal of hidden state and multi-scale consistency loss, respectively. TMEM denotes the transformer-based motion embedding module and DA is the proposed data augmentation strategy. It can be seen that the removal of these components leads to different degrees of visual degradation in the recovered GS image, such as blurring artifacts and missing details at the pant legs and human faces. In contrast, our full model can reconstruct a higher-quality GS image faithfully, where sharper and richer image details are retained.

3. Limitation and Discussion

The RSC model is trained based on a specific readout time ratio [16], which makes it less effective to generalize to RS data with large readout bias. This is a common problem for learning-based RSC methods [3,4,10,15], as pointed out in [3, 5, 14]; while traditional RSC methods [6, 12, 16, 17] rely on non-trivial readout calibration to solve it. The development of RSC methods that are robust to readout time ratios will be an attractive future topic. In addition, our method may suffer from inaccurate local details when encountering low-texture or narrow objects. We show visual results of two sets of failure cases in Fig. S2, where artifacts appear in the thin white pillars and the rear seam of the car. We reckon this is because motion estimation is relatively more difficult in these challenging narrow image regions, which may easily lead to context misalignment. The use of adaptive warping [1–3] or multiple motion fields strategy [3, 7] may alleviate this problem.

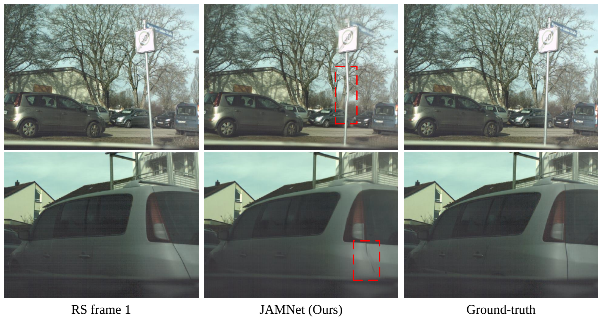


Figure S2. Failure cases in challenging narrow image regions.

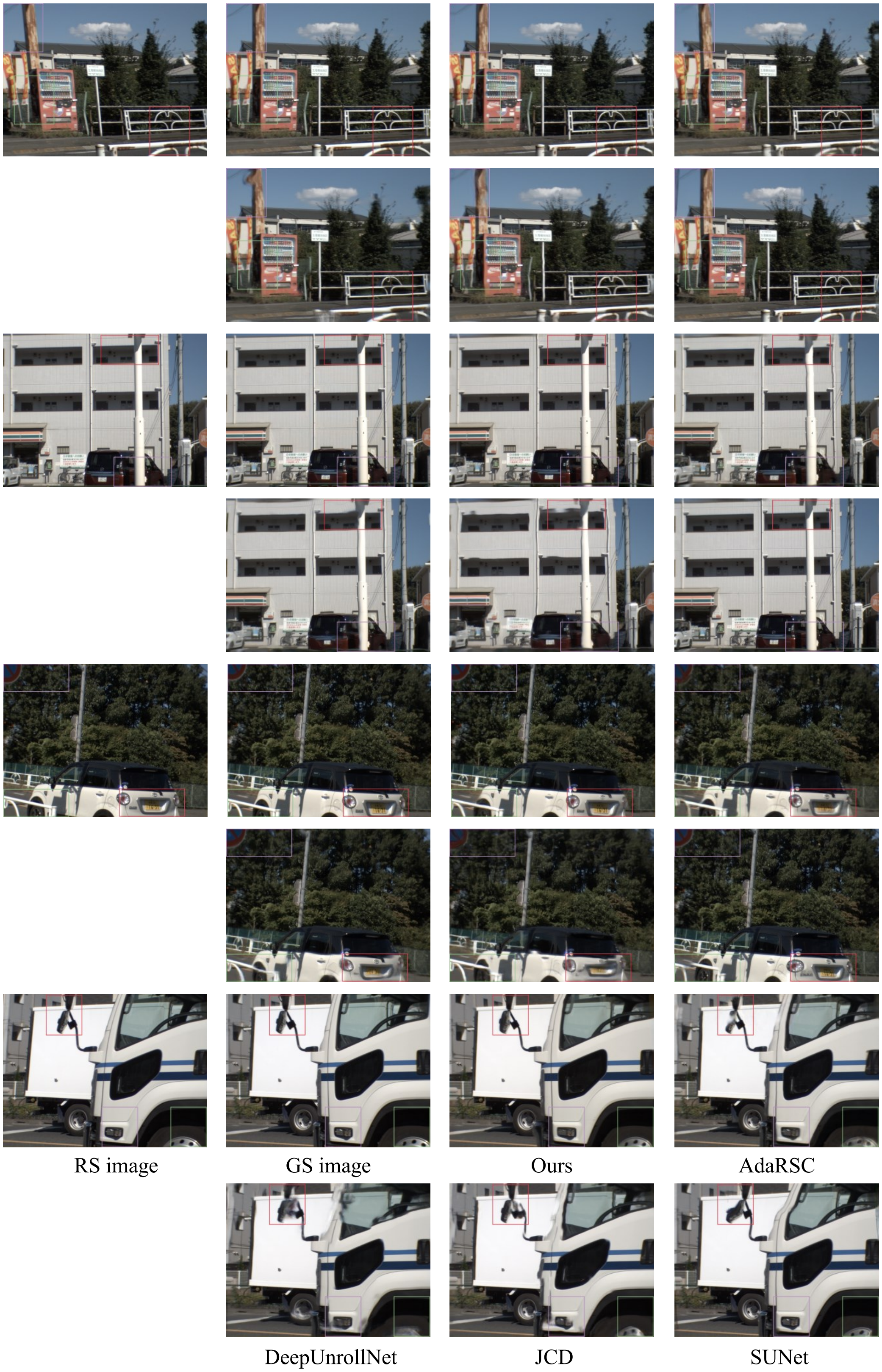


Figure S3. **More real-world RSC results on the BS-RSC dataset [3].** Existing RSC methods either fail to remove the RS effect or introduce other undesirable artifacts (*e.g.*, blurring, ghosting, unsmoothing, missing details, local errors, *etc*.). In contrast, thanks to the coarse-to-fine joint appearance and motion learning, our method achieves higher-quality results. Best viewed on screen.

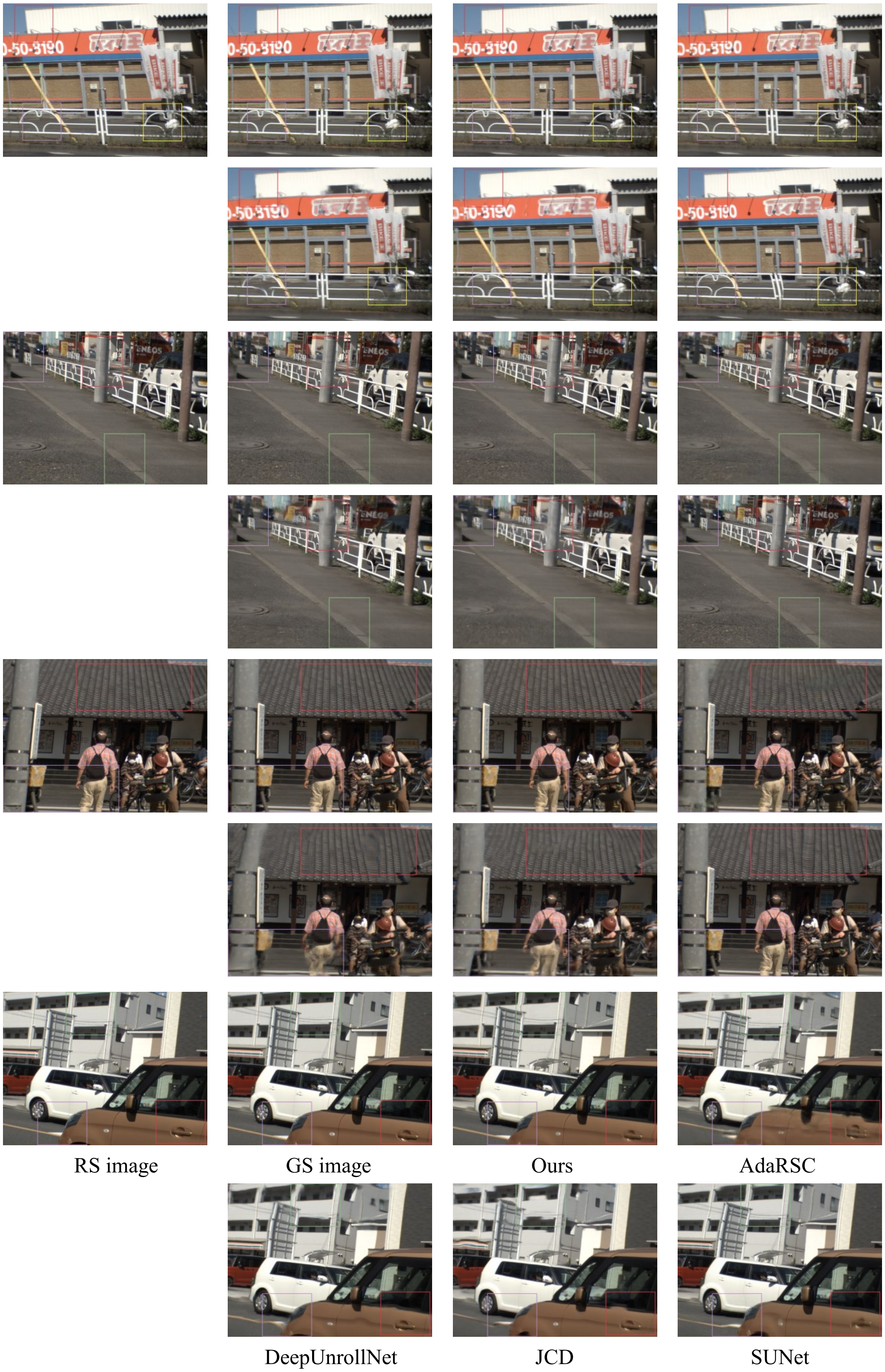
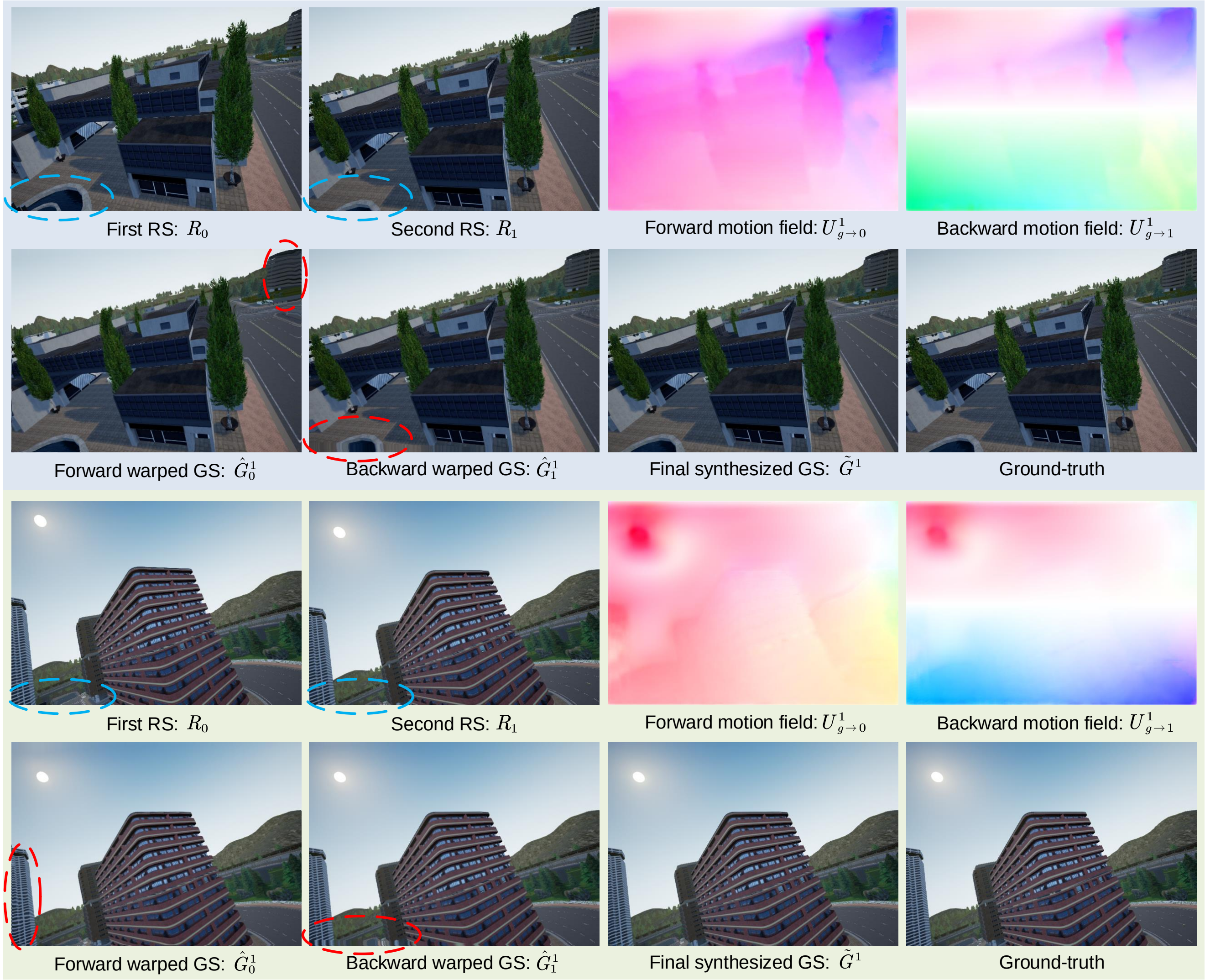


Figure S4. **More real-world RSC results on the BS-RSC dataset [3].** Existing RSC methods either fail to remove the RS effect or introduce other undesirable artifacts (*e.g.*, blurring, ghosting, unsmoothing, missing details, local errors, *etc.*). In contrast, thanks to the coarse-to-fine joint appearance and motion learning, our method achieves higher-quality results. Best viewed on screen.



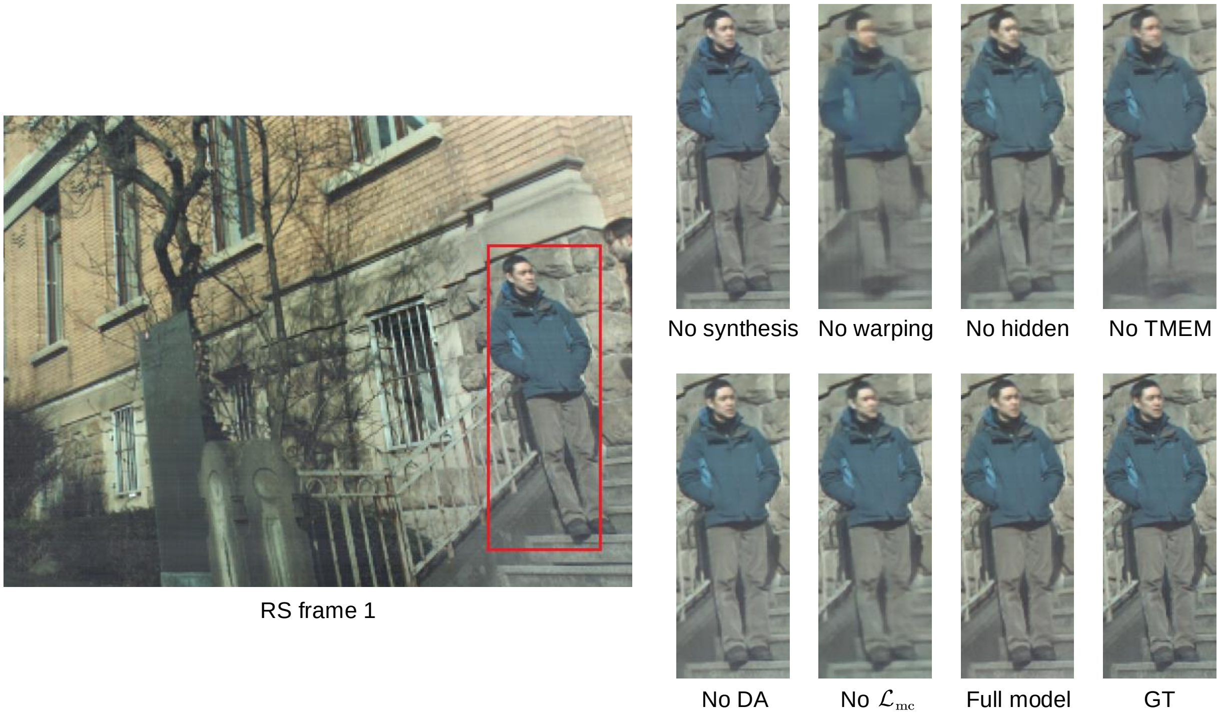
First RS: R_0 Second RS: R_1 Forward motion field: $U^1_{g\rightarrow 0}$ Backward motion field: $U^1_{g\rightarrow 1}$

Forward warped GS: $\hat{G}^1_0$ Backward warped GS: $\hat{G}^1_1$ Final synthesized GS: $\tilde{G}^1$ Ground-truth

First RS: R_0 Second RS: R_1 Forward motion field: $U^1_{g\rightarrow 0}$ Backward motion field: $U^1_{g\rightarrow 1}$

Forward warped GS: $\hat{G}^1_0$ Backward warped GS: $\hat{G}^1_1$ Final synthesized GS: $\tilde{G}^1$ Ground-truth

Figure S5. **Visual results of the intermediate process of our method.** Due to the occlusion between the RS and GS images (*cf*. blue circles), both forward and backward warped GS images suffer from local detail loss, as shown by the red circles. Fortunately, our method can aggregate these contextual information in an adaptive manner to synthesize a final high-fidelity GS image. Furthermore, our method can recover a plausible bilateral undistortion field with significant scanline-dependent properties (*e.g.*, the upper and lower regions of $U^1_{g\rightarrow 1}$ exhibit different pixel displacement directions).


RS frame 1

No synthesis No warping No hidden No TMEM

No DA No $\mathcal{L}_{\mathrm{mc}}$ Full model GT

Figure S6. **Visualization of ablation results.** Here, TMEM indicates the transformer-based motion embedding module, and DA denotes the proposed data augmentation strategy. See Table 3 in the main text for the corresponding quantitative results. Overall, our full model recovers the most visually appealing GS images with fewer local artifacts and sharper edges, such as pant legs and human faces.

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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3703–3712, 2019. 2

[2] Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. MEMC-Net: motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):933–948, 2021. 2

[3] Mingdeng Cao, Zhihang Zhong, Jiahao Wang, Yinqiang Zheng, and Yujiu Yang. Learning adaptive warping for real-world rolling shutter correction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17785–17793, 2022. 1, 2, 3, 4

[4] Bin Fan, Yuchao Dai, and Mingyi He. SUNet: symmetric undistortion network for rolling shutter correction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4541–4550, 2021. 1, 2

[5] Bin Fan, Yuchao Dai, and Hongdong Li. Rolling shutter inversion: bring rolling shutter images to high framerate global shutter video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

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

[7] Ping Hu, Simon Niklaus, Stan Sclaroff, and Kate Saenko. Many-to-many splatting for efficient video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3553–3562, 2022. 2

[8] 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)*, pages 624–642. Springer, 2022. 1

[9] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5754–5763, 2019. 2

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

[11] Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, and Jiaya Jia. Video frame interpolation with transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3532–3542, 2022. 1

[12] Subeesh Vasu, Mahesh MR Mohan, and AN Rajagopalan. Occlusion-aware rolling shutter rectification of 3d scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 636–645, 2018. 2

[13] Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I Chang, Yan Xu, et al. MaskFlowNet: asymmetric feature matching with learnable occlusion mask. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6278–6287, 2020. 2

[14] Zhihang Zhong, Mingdeng Cao, Xiao Sun, Zhirong Wu, Zhongyi Zhou, Yinqiang Zheng, Stephen Lin, and Imari Sato. Bringing rolling shutter images alive with dual reversed distortion. *arXiv preprint arXiv:2203.06451*, 2022. 2

[15] Zhihang Zhong, Yinqiang Zheng, and Imari Sato. Towards rolling shutter correction and deblurring in dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9219–9228, 2021. 1, 2

[16] Bingbing Zhuang, Loong-Fah Cheong, and Gim Hee Lee. Rolling-shutter-aware differential sfm and image rectification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 948–956, 2017. 2

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