Supplementary Material for Learning Adaptive Warping for Real-World Rolling Shutter Correction

Mingdeng Cao[1] Zhihang Zhong[2] Jiahao Wang[1] Yinqiang Zheng[2] ✉ Yujiu Yang[1] ✉
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]The University of Tokyo

In this supplementary document, we provide additional materials to supplement our main submission. Specifically, in Sec.1, we further provide more analysis of multiple motion field prediction strategy. Sec.2 additionally provides more details of the encoder network and decoder network in the proposed method. Sec.3 gives more visual results on the proposed real-world dataset BS-RSC and synthesized dataset Fastec-RS [2] to demonstrate the effectiveness of the proposed method.

1. Analysis of Multiple Motion Fields Strategy

We provide an ablation study of the number of the motion fields used by our model in our main paper, and multiple motion fields can improve the correction performance. As a result, the proposed multiple motion fields strategy can alleviate the inaccurate estimation problem. The diverse fields mean that more information can be aggregated during the warping process, providing alternative reasonable motions to alleviate artifacts in the warped frame. Here, we further did some explorations to validate the generality of this strategy on the backward warping-based model. Fig. 1 shows the variation of PSNR and diversity (measured by the standard deviation) w.r.t. the number of motion fields. Obviously, both PSNR and STD grow rapidly when increasing fields, and then reach the peak at 9 groups, followed by a slight drop when further increasing the number of motion fields. This means the diversity of the predicted fields plays a significant role in the warping process for high-quality RSC.

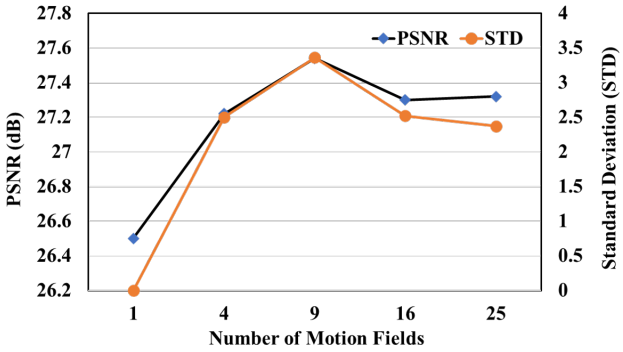


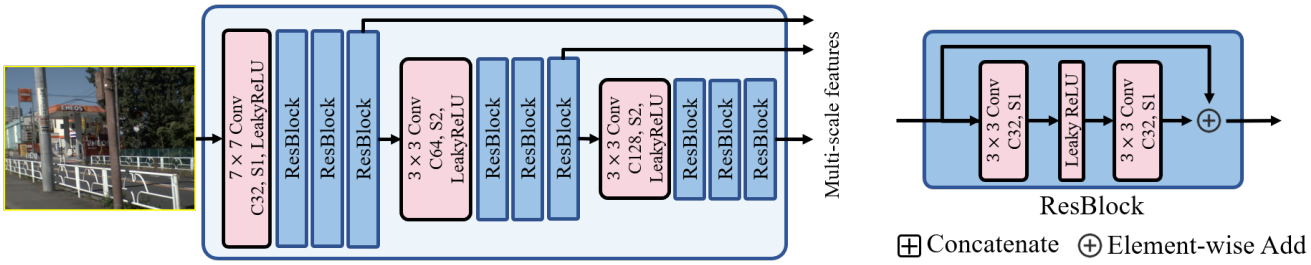
Figure 1. The PSNR and deviation of the motion fields w.r.t. the groups of motion fields.

2. Network Details

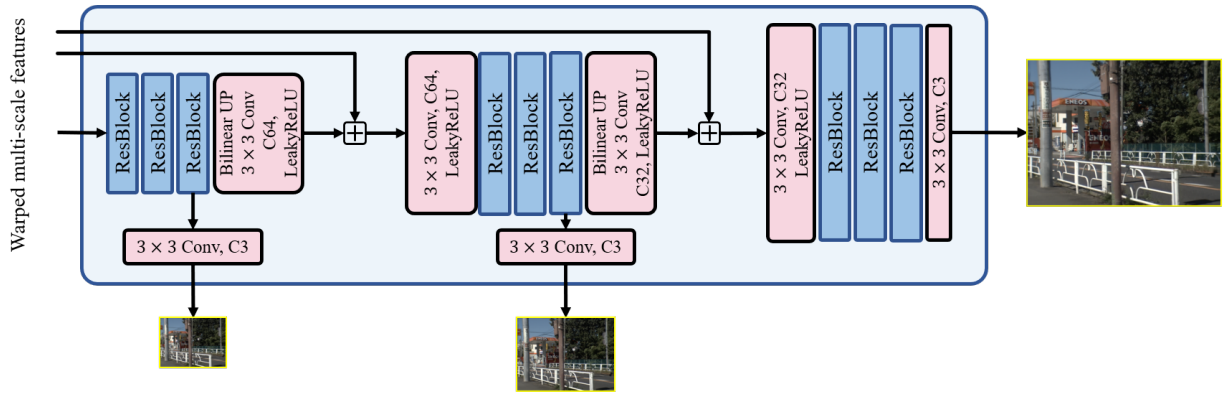
In this section, we present details of the frame-level feature extraction network and coarse-to-fine decoder network in Sec. 3.2 in our main submission.

Feature Extraction Network. The details of the feature extraction network is shown in Fig. 2(a), which consists of three stages in a pyramid style. Three consecutive RS frames are fed into the feature extractor, and outputs multi-scale features corresponding to each RS frame. The first scale utilize a large convolution kernel 7×7 to translate the image into feature maps, then three residual blocks [1] are adopted to extract features. As for the second and third stage, we adopt a convolution layer with stride 2 to downsample the feature maps and double the number of the channels. The extracted multi-scale features are then warped by the following adaptive warping module.

Decoder Network. Fig. 2(b) illustrates the details of the decoder network's architecture, which reconstructs the GS frame in a coarse-to-fine manner. The multi-resolution corrected GS frames are used to compute the multi-scale loss. The decoder's structure is similar to the encoder's, and three stages are adopted. For each stage, we first fuse the concatenated features of the upsampled features from the previous stage and the features from the encoder network with a convolution layer. Then three residual blocks refine the features and output the corrected GS frame at the current scale. These features are last upsampled by bilinear interpolation followed by a convolution layer.



(a) Feature extractor



(b) Decoder Network

Figure 2. **(a)** The details of frame-level feature extraction network, which extract multi-scale features from each RS frame. **(b)** The details of decoder network. The warped multi-scale features are fed into model and output multiple GS frames with different resolutions.

3. Additional Qualitative Experimental Results

We present additional experimental results to demonstrate the excellent performance of the proposed methods qualitatively. We first provide more visual comparisons on the synthesized dataset Fastec-RS Fig. 3, where the proposed method shows high competitive performance. As for the real RS distortions, Fig. 4 and Fig. 5 shows the experimental results on the proposed real-world RS distorted dataset BS-RSC. Our model can remove these real distortions better. These experimental results verify the effectiveness of our model to remove the RS distortions and generate high-quality potential GS frames. We also provide a video demo "video_demo.mp4" in the uploaded supplementary materials to show the visual comparison and temporal consistency.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[2] Peidong Liu, Zhaopeng Cui, Viktor Larsson, and Marc Pollefeys. Deep shutter unrolling network. In *CVPR*, pages 5941–5949, 2020. 1, 3, 4, 5

[3] Bingbing Zhuang, Loong-Fah Cheong, and Gim Hee Lee. Rolling-shutter-aware differential sfm and image rectification. In *CVPR*, pages 948–956, 2017. 3, 4, 5

(a) RS frame (b) Zhuang *et al*. [3] (c) DSUN [2]

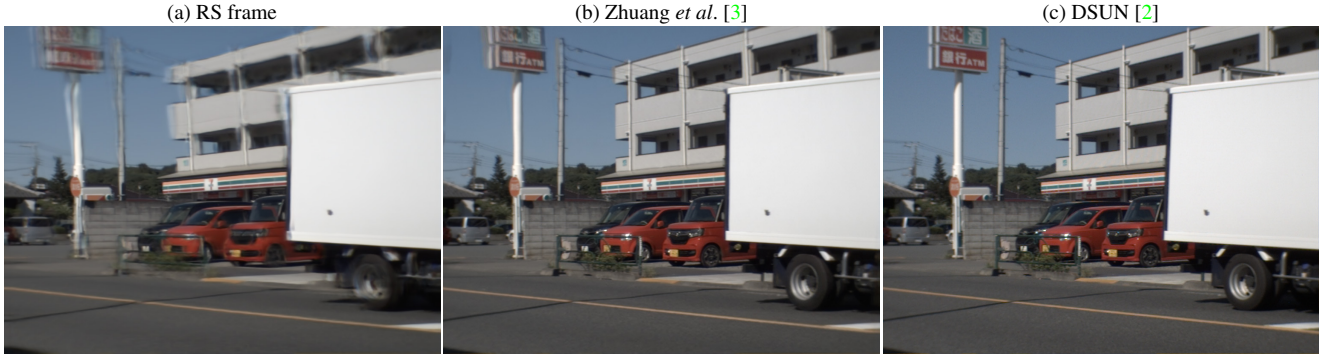
(e) JCD (f) Ours (g) GS frame

Figure 3. More visual comparisons on the synthesized Fastec-RS dataset.

(a) RS frame (b) Zhuang *et al*. [3] (c) DSUN [2]

(e) JCD (f) Ours (g) GS frame

(a) RS frame (b) Zhuang *et al*. [3] (c) DSUN [2]



(e) JCD (f) Ours (g) GS frame

Figure 4. More visual comparions on the proposed BS-RSC dataset.

(a) RS frame (b) Zhuang *et al*. [3] (c) DSUN [2]

(e) JCD (f) Ours (g) GS frame

(a) RS frame (b) Zhuang *et al*. [3] (c) DSUN [2]

(e) JCD (f) Ours (g) GS frame

Figure 5. More visual comparions on the proposed BS-RSC dataset.