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

Abstract

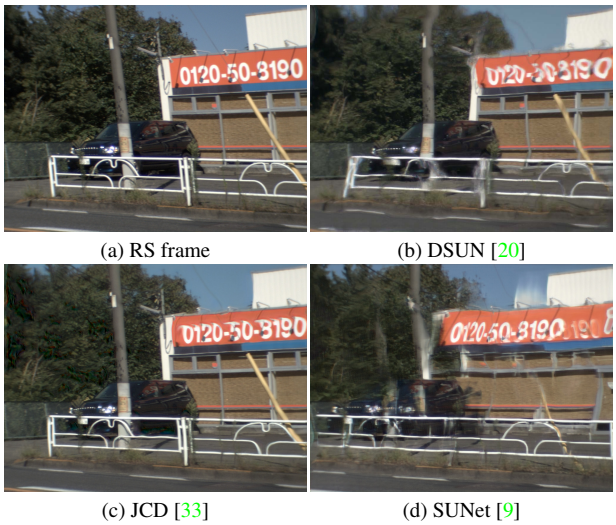
This paper proposes the first real-world rolling shutter (RS) correction dataset, BS-RSC, and a corresponding model to correct the RS frames in a distorted video. Mobile devices in the consumer market with CMOS-based sensors for video capture often result in rolling shutter effects when relative movements occur during the video acquisition process, calling for RS effect removal techniques. However, current state-of-the-art RS correction methods often fail to remove RS effects in real scenarios since the motions are various and hard to model. To address this issue, we propose a real-world RS correction dataset BS-RSC. Real distorted videos with corresponding ground truth are recorded simultaneously via a well-designed beam-splitter-based acquisition system. BS-RSC contains various motions of both camera and objects in dynamic scenes. Further, an RS correction model with adaptive warping is proposed. Our model can warp the learned RS features into global shutter counterparts adaptively with predicted multiple displacement fields. These warped features are aggregated and then reconstructed into high-quality global shutter frames in a coarse-to-fine strategy. Experimental results demonstrate the effectiveness of the proposed method, and our dataset can improve the model's ability to remove the RS effects in the real world. The project is available at https://github.com/ljzycmd/BSRSC.

1. Introduction

Most consumer cameras adopt CMOS sensors for imaging due to their low power consumption, compact design, and fast imaging. At the same time, most CMOS sensors have rolling shutter (RS) effects during imaging. Unlike a global shutter (GS) camera capturing all pixels simultaneously, an RS camera sequentially captures the image pixels row by row. Therefore, the RS distortions would occur in the recorded images and videos when relative movements arise between the camera and objects. The RS distortions significantly impair the visual quality. Moreover, the distorted images and videos deteriorate the performance of some downstream tasks, like 3D reconstruction, pose estimation, and depth prediction [3, 8, 10, 16], leading to erroneous, undesirable, and distorted results.

There are usually two ways to mitigate the performance gap of existing computer vision algorithms working on the RS distorted and GS images. The first is to keep the original RS images unchanged and adapt the algorithms to the RS distorted images. Thus, many RS-aware algorithms are proposed in 3D vision field, *e.g.*, RS structure-from-motion reconstruction [13, 34], RS stereo [27], RS camera calibration [22] and RS absolute camera pose [1, 3, 4, 18]. An arguable better way is to correct the RS distorted images into GS images. In this way, we don't need to modify existing vision algorithms and can obtain visual-friendly images. Therefore, correcting the rolling shutter (RSC) images is increasingly becoming significant in photography and has attracted considerable research attention recently [2, 9, 20, 24].

Existing RS effect removal methods can be categorized into single-image- and multi-frame-based. When restoring the GS image from only one RS image, many external constraints or priors (*e.g.*, geometric priors) are adopted [17, 24, 25, 35] since it is a highly ill-posed problem. Compared to single-image-based correction, multi-frame-based methods are more general and can utilize motion information for correction. Due to the great success of convolutional neural networks (CNNs) on various computer vision tasks and the proposed synthesized RSC datasets, researchers designed specific model architectures to remove the RS distortions in an end-to-end manner based on multiple frames. Usually, the motions across multiple frames are modeled first. Then the GS image corresponding to the reference RS frame is restored by warping operations. For instance, Liu *et al*. [20] predict velocity field from the correlation volume, and Fan *et al*. [9] utilize PWC-Net framework [28] to estimate the undistortion flow to correct the RS frame. They both adopt forward warping to remove the RS effect, and have achieved some promising results. However, the corrected GS images still suffer from blurs and texture detail loss for the following reasons: 1) The modeled motions are inaccurate since there is no ground truth for supervision dur-



(a) RS frame (b) DSUN [20]

(c) JCD [33] (d) SUNet [9]

Figure 1. The real-world rolling shutter correction results of existing state-of-the-art methods trained with synthesized data. We see that all methods failed to remove the RS effects and even introduced many artifacts into the corrected frame.

ing the training process. 2) The warping operations are not learnable, which cannot aggregate the features adaptively. 3) Meanwhile, some regions in the potential GS frame do not appear in the input RS frames. Thereby, it is difficult for the model to generate unseen areas. 4) Moreover, these models are trained on the synthesized RSC datasets where the motions are rather monotonous. And many artifacts exist in the synthesized RS frames, greatly restricting the model's performance on the natural RS image correction. Fig. 1 shows some real-world RSC results of state-of-the-art methods trained with synthesized data.

To move beyond these limitations mentioned above, we propose a novel adaptive warping module and a real-world dataset for rolling shutter correction. Our model takes three consecutive frames as input, restoring the GS frame corresponding to the central RS frame at intermediate imaging time. We propose an adaptive warping module to better exploit high-quality GS frame restoration by mitigating inaccurate RS motion estimation and warping problems. Firstly, multi-scale features of each RS frame are extracted. Then, we construct a correlation volume to build the correspondence between central and neighboring RS features. The volume is used to predict multiple motion fields rather than only one generated in previous works [9, 20]. After that, an adaptive attention mechanism is proposed to warp the RS features by aggregating the contextual features according to the predicted motion fields. The designed warping process is learnable, aggregating the features to the GS-aware features attentively and adaptively. Note that we perform adaptive warping at all scales. A decoder network further decodes these warped multi-scale features and reconstructs the corresponding GS frame. The proposed model can be trained in an end-to-end manner.

Considering the performance gap on the synthesized datasets and real RS distorted scenarios, we propose BS-RSC, the first dataset for real-world RSC with various motions in dynamic scenes, collected by a well-designed beam-splitter acquisition system. An RS camera and a GS camera are physically aligned to capture RS distorted and GS frames simultaneously.

Our contributions can be summarized as follows:

- We propose a novel feature warping module for rolling shutter correction, which adaptively warps RS features into global counterparts for high-quality GS frame restoration.
- We contribute BS-RSC, the first real-world RSC dataset (devoid of motion blur) with various motions collected by a well-designed beam-splitter acquisition system, bridging the gap for real-world RSC task.
- The quantitative and qualitative experimental results on real-world and synthetic datasets show the excellent performance of the proposed method against the state-of-the-art methods.

2. Related Works

2.1. Deep Rolling Shutter Correction

CNNs are used to remove the RS effects due to the considerable success in many computer vision tasks. For single image RSC, Rengarajan *et al*. [24] proposed a CNN architecture to estimate the row-wise camera motion from a single image and undo RS distortions back to the time of the first-row exposure. They adopted a long rectangular convolutional kernel to learn the effects produced by row-wise exposure specifically. Zhuang *et al*. [35] further proposed a structure-and-motion-aware RS correction model that reasons about the concealed motions between the scanlines as well as the scene structure, where the camera scanline velocity and depth are estimated.

Since single image RSC is a highly ill-posed task, multi-frame RSC can perform better by modeling the RS motion more accurately and has recently received much attention. Liu *et al*. [20] proposed an end-to-end network for RSC by predicting dense displacement field from two consecutive RS frames. Then they adopted a differentiable forward warping module to warp the RS image into the global one. Further considering the blurs in the RS distorted images, Zhong *et al*. [33] proposed the first real-world rolling shutter correction and deblurring (RSCD) and a joint correction and deblurring (JCD) model to tackle the the RSCD problem. Most recently, Fan *et al*. [9] utilized PWC-Net [28] to

predict symmetric undistortion fields and restore the potential GS frames by a time-centered GS image decoder network, achieving promising results on the synthetic datasets. These methods still suffer from the blurs and detail loss in the restored GS frame due to the inaccurate displacement field estimation and warping. To alleviate such artifacts, we propose to predict multiple fields and warp the RS features adaptively.

2.2. Attention Mechanism

Attention mechanism was introduced [5] for machine translation, and has been widely used in both natural language processing and computer vision. In [29], a novel Transformer architecture was constructed using attention as a primary mechanism, and it replaced the recurrent structure with the self-attention operation. Thanks to the powerful long-range and relation modeling capacities of attention, it was gradually introduced to vision tasks and has achieved considerable success [14, 23, 30].

Recently, attention mechanism or Transformer has been adapted to image or video restoration tasks and achieved great success, *e.g.*, super resolution [7, 19, 31]. In [31], the authors proposed a texture transformer network for reference-based image super-resolution, which adopts an attention mechanism to transfer the texture details from the reference image adaptively. Chen *et al*. [7] proposed an Image Process Transformer (IPT) for various image restoration tasks, *e.g.*, super-resolution, denoise, by task-specific heads and tails. Liang *et al*. [19] utilized Swin Transformer [21] for multiple image restoration tasks and achieved better performance with much fewer parameters. Attention has shown high potential for vision tasks. This paper also exploits the attention mechanism for adaptive warping to restore high-quality GS frames.

2.3. RSC Dataset Synthesis

Note that CNN-based approaches usually require a large amount of training data to learn the correction from RS to GS image. However, current RSC data or publicly available datasets are synthesized, where the RS images are generated from the captured GS images. For example, in [24], an affine transformation corresponding to RS motions is used to synthesize RS images. Zhuang *et al*. [35] synthesized RS images by warping a single GS image from KITTI dataset [11] with dense depth map and camera motions. In [2], various simulated motions are used to generate RS images. Recently, researchers in [20] proposed two datasets, Carla-RS and Fastec-RS datasets, which generate more realistic RS distorted images via high-speed cameras and simulate the natural RS image formation process beyond camera motions or 3D geometry. The Carla-RS is synthesized from a free-moving rolling shutter camera in a virtual 3D Carla simulator. On the contrary, the Fastec-RS dataset is created using the GS images in the real world with a 2400 FPS global shutter camera. However, the synthesized RS images are unnatural and full of line artifacts (shown in Fig. 5). Moreover, most of the scenes in Fastec-RS are collected by a horizontally moving camera, while various motions cause the RS images in the real world. These limitations significantly deteriorate the performance of RSC models. This paper proposes the first real-world RSC dataset for model training to restore high-quality GS images from real-world RS distorted images.

3. Proposed Method

3.1. Problem Formulation

As described in [20], the GS frame can be restored by warping the RS features backward with predicted displacement filed:

$$\mathbf{I}^g(x) = \mathbf{I}^r(x + \mathbf{U}_{g\to r}(x)), \tag{1}$$

where $\mathbf{I}^g$ is the potential GS frame; $\mathbf{I}^r$ is the input RS frame; $\mathbf{U}_{g\to r}$ is the displacement field from GS to RS frame, and x is a certain pixel. It is difficult to estimate the displacement field $\mathbf{U}_{g\to r}$ since only the RS frames are available. Fortunately, the velocity vector can be estimated from the optical flow $\mathbf{V}$ between two consecutive RS frames. Thus the displacement can be calculated when the velocity is constant:

$$\mathbf{U}(x) = \lambda \mathbf{V}(x)\mathbf{T}(x), \tag{2}$$

where λ is a scaling factor, and $\mathbf{T}(x)$ is the time offset corresponding to the middle scanline of RS frame. Therefore, existing methods try to estimate the displacement filed firstly from two consecutive RS frames, then warp the RS features with a differentiable forward warping block (DFW) [20]. A DFW module attempts to approximate the intensity of a particular pixel x in the potential GS image by aggregating its neighboring pixel intensities in the RS features with weights proportional to the neighbor's distance, *i.e.*, the greater the distance of a neighbor, the smaller its weight. Therefore, accurate motion estimation and warping are two key factors to restore the potential GS frames. However, the accurate $\mathbf{U}$ is hard to estimate since $\mathbf{U}$ cannot be effectively supervised during training when only the GS frame is adopted for supervision. The inaccurate estimation $\mathbf{U}$ further results in undesired warping results since the DFW module aggregates the neighboring pixels to x with distance-aware weights. As a result, the corrected GS frame often suffers from blurs and other artifacts.

3.2. Model Overview

Our model aims to alleviate inaccurate displacement field estimation and error-prone warping problems with multiple fields prediction and adaptive warping module.

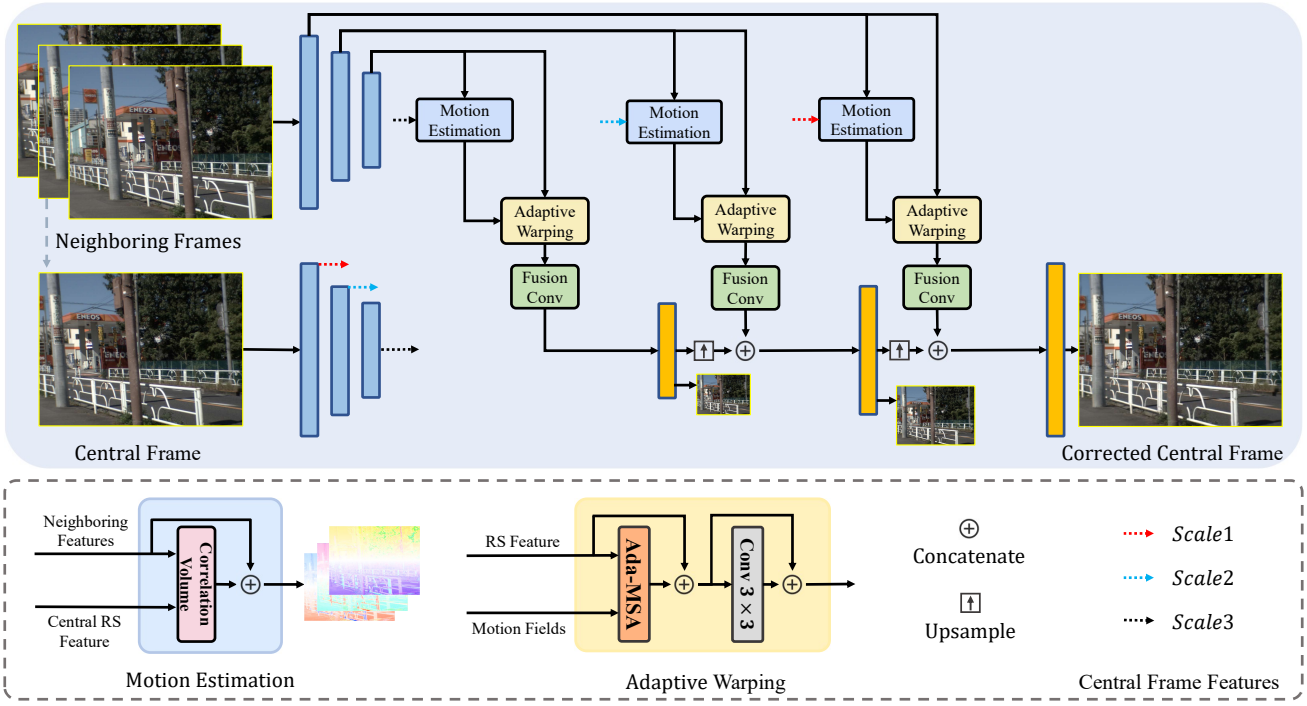


Figure 2. Main architecture of the proposed RSC model. Our model tries to predicts multiple displacement fields rather than only one to alleviate existing inaccurate motion estimation. We also propose an adaptive warping module to warp the RS features into the GS one adaptively under the guidance of the bundle of fields.

Building on current CNN-based RSC methods, our model inputs three consecutive RS frames to explore motion information and complementary contextual information, and restore the GS frame at the intermediate exposure time (middle scanline) of the input central RS frame. Our model consists of three parts shown in Fig. 2: a multi-scale feature extractor, an adaptive warping module, and a coarse-to-fine GS frame decoder. We first extract frame-level multi-scale features. Then, for the features at each scale, the neighboring RS features are used to predict the forward and backward motion information and warped by the proposed adaptive warping module. These warped features are fused by a convolution block. Last, the decoder decodes the warped features and outputs the corrected GS frame in a coarse-to-fine manner.

3.3. Adaptive Warping Module

Multiple Displacement Fields Generation. A key difference from previous methods is that our model predicts multiple displacement fields rather than one for warping. Moreover, the constant velocity assumption is too restrictive in Eq. (2), thus we modulated the multiple displacement fields by further predicting weights. Specifically, for the t-th RS feature $F_t^l \in \mathbb{R}^{C\times H\times W}$ at l-th scale, we first construct a 3D correlation volume $\mathbf{CV_t^l}$ [28] to build the correspondence with central RS frames. Then the volume is used to predict multiple displacement fields and their weights by a residual block [12]:

$$\{\mathbf{U}_t^{l,0},\ldots,\mathbf{U}_t^{l,M-1},\mathbf{W}\} = \text{ResBlock}([\mathbf{CV}_t^l, F_t^l]), \quad (3)$$

where l index the scale, and M is the number of motion fields. Each field contains two channels corresponding to the horizontal and vertical movements. $\mathbf{W} \in \mathbb{R}^{M\times H\times W}$ is the weight of each estimated field. So the final predicted displacement fields are modulated by multiplying the estimated weights.

Adaptive Warping. As for the warping process, we proposed an Adaptive Warping Module (AWM) utilizing self-attention to aggregate the features sampled under the predicted multiple displacement fields. AWM consists of an adaptive multi-head attention (Ada-MSA) and a convolutional block. The Ada-MSA mechanism is shown in Fig. 3. Firstly, for each pixel x (consists of row index i and column index j) in t-th RS features F_t^l at scale l, the query vector Q is generated by a linear transformation with matrix W_q:

$$Q = W_q F_l^t(x). \quad (4)$$

Subsequently, the feature set $N(x)$ are sampled under the guidance of estimated multiple displacement fields $\mathbf{U}_t^l$:

$$N_t^l(x) = \{F_t^l(x + \mathbf{U}_t^{l,i}(x)) | i = 0, 1, \ldots, M-1\}. \quad (5)$$

Then the key K and value V vectors are then generated by a linear transformation from the sampled features:

$$K = W_k N_t^l(x), V = W_v N_t^l(x), \quad (6)$$

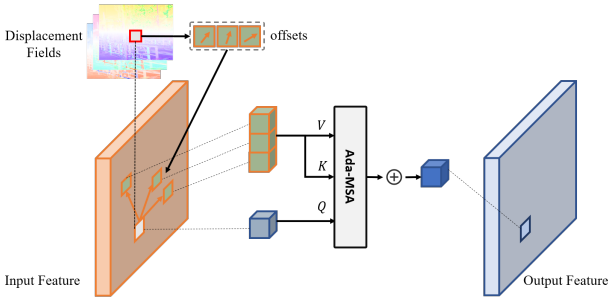


Figure 3. Illustration of adaptive multi-head self-attention mechanism (Ada-MSA). Ada-MSA aims to warp the input RS features into the GS features adaptively under the guidance of estimated multiple motion fields.

where $W_k \in \mathbb{R}^{d\times C}$ and $W_v \in \mathbb{R}^{d\times C}$ are transformation matrices. Thus the adaptive attention feature at h-th head is calculated by

$$\text{AdaMSA}_h(x, F_t^l, \mathbf{U}_t^l) = \text{SoftMax}(\frac{Q_h^T K_h}{\sqrt{d_h}})V_h^T, \quad (7)$$

where h indexes the attention head, and Q_h, K_h and V_h are with dim $d_h = \frac{d}{H}$. The outputs of all H heads are concatenated into d dims vector and projected to the output feature. Through this adaptive warping module with multiple multiple motion fields, the RS features are aggregated to the GS counterpart adaptively.

3.4. Loss Functions

We train the proposed model in an end-to-end manner, and only the ground truth GS frame is required for supervision. Following previous work [33], we adopt the Charbonnier loss $\mathcal{L}_c$ and perceptual loss $\mathcal{L}_p$ to ensure the visual quality of the corrected GS frame. The total variation loss $\mathcal{L}_{tv}$ is adopted to ensure the smoothness in the estimated displacement field. Thus total loss can be formulated as:

$$\mathcal{L} = \mathcal{L}_c + \lambda_p \mathcal{L}_p + \lambda_{tv} \mathcal{L}_{tv}. \quad (8)$$

4. BS-RSC Dataset

A real-world dataset without synthetic artifacts is essential to improve the capacity of real applications of CNN-based RSC methods. Recently, some specific optical acquisition systems have been designed to capture the real-world image or video pairs for restoration tasks, improving the generalization capacity of CNN models. Cai *et al*. [6] constructed a real-world super-resolution dataset where paired high- and low-resolution data of the same scene are captured by adjusting the focal length of a digital camera. For deblurring, Rim *et al*. [26] and Zhong *et al*. [32] collected real-world single image and video deblurring dataset respectively, adopting a beam-splitter acquisition system. Inspired by these pioneering works, we also propose a beam-splitter acquisition system to collect the first real-world dataset for the RSC task, termed as BS-RSC.

4.1. Beam-Splitter Acquisition System

The architecture of the designed beam-splitter acquisition system is shown in Fig. 4(a), where a beam-splitter splits the incoming light into two beams and passes them into the following RS and GS cameras. We choose the FLIR FL3-U3-13S2C RS camera with a 1/3-inch CMOS sensor (3.63 um pitch size) and the FLIR GS3-U3-28S4C GS camera with a 1/1.8-inch CCD sensor (3.69 μm pitch size). These two cameras are geometrically aligned via the 50/50 beam splitter. With the aid of a laser beam, we first adjust the alignment mechanically towards an accuracy of a few pixels. After that, we conduct a homography correction with a standard checker pattern to further reduce misalignment to subpixel level. The exposure time of both the RS and the GS camera is 1ms, avoiding blurs in the captured video. Both cameras run at 25 fps. We use a wave generator to generate synchronized pulses at 25Hz, and the phase of the pulse for the GS camera is properly delayed, such that the GS exposure timestamp matches the middle scanline of the RS camera (shown in Fig. 4(b)). As for photometric alignment, we put a neutral density filter before the RS camera to equalize the sensitivity of the two cameras. We further use a color checker pattern to correct the RGB response of the GS camera, such that both cameras share the same color response. The whole system is just about one kilogram, thus can be held easily and moved freely.

4.2. Data Composition

The collected BS-RSC contains RS videos with various camera and object motions, mainly in outdoor street scenes with cars and people, *etc*. Specifically, the designed beam-splitter acquisition system collects a total of 80 RS-GS HD (1024 × 768) video pairs, and each video contains 50 frames. We further divide it into Train set, Val set, and Test set with 50, 15, 15 videos, respectively.

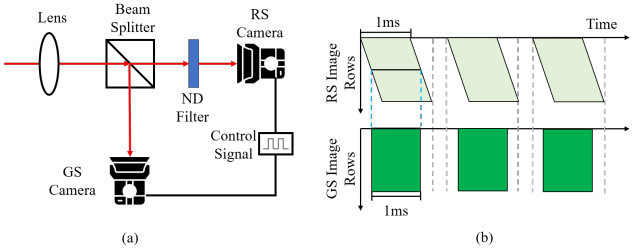


Figure 4. The designed beam-splitter acquisition system for real-world RSC dataset construction. **(a)** structure of the designed beam-splitter acquisition system. **(b)** exposure scheme of the GS and RS camera. The acquisition system can capture the GS frame at the intermediate exposure time of RS frame.

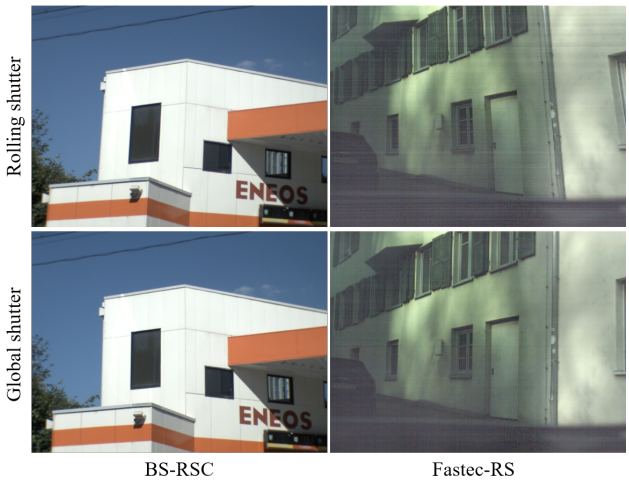


Figure 5. **Left**: The real world RS-GS example in the collected BS-RSC dataset. **Right**: The synthesized RS-GS example in the Fastec-RS dataset [20]. We see that our real RS frame is more natural, and there are much artifacts in the synthesized RS frames.

5. Experiments

5.1. Experimental Setting

Datasets. We conduct experiments on the proposed real-world BS-RSC dataset. Besides, we also provide experimental results on the popular synthesized dataset Fastec-RS [20], which contains 76 video sequences, and each video contains 34 frames. Note that we use the dataset coming from the public released dataset, which is slightly different from the description in the original paper. Both test and validation subsets are used to calculate the metrics.

Implementation Details. During training, three consecutive RS frames in RGB style are fed into our model. The input frames are first randomly cropped into 480×256 and randomly flipped horizontally for data augmentation. λ_p and $\lambda_t v$ are set to 0.01 and 0.001, respectively. The initial learning rate is set to 2×10^{-4}, and the ADAM [15] is adopted to optimize the model parameters. The model is trained for 400 epochs with a cosine annealing learning rate adjusting scheduler. For testing, three consecutive frames are fed into the model directly without any augmentation. We set the number of displacement fields $M = 9$ in the following experiments.

Evaluation Metrics and Methods of Comparison. Both PSNR and SSIM are employed to evaluate the corrected results quantitatively. Visualizations of the corrected RS frames are shown for qualitative comparison. We compare the proposed method to state-of-the-art RSC method, including a traditional methods proposed in [34], CNN-based methods DSUN [20], JCD [33] and SUNet [9]. These methods have shown promising effectiveness on the synthesized dataset. As the authors of SUNet have not yet published the code and test results, we cannot report any results other than those in the original paper.

5.2. Comparison to the State-of-the-art.

Results on BS-RSC. The quantitative comparison of the proposed real-world dataset BS-RSC is shown in Tab. 1. Thanks to the multiple motion fields prediction and the adaptive warping strategy, our model achieves the best PSNR and SSIM evaluation metrics with a large performance improvement than SOTA methods. The qualitative comparison is shown in Fig. 6. We see that the proposed method obtains more visually friendly results than other methods (*e.g.*, the billboard and the trees). These superior performances significantly demonstrate the effectiveness of our model on real-world rolling shutter correction.

Methods	PSNR↑(dB)	SSIM↑
Zhuang *et al*. [35]	19.80	0.698
DeepUnrollNet [20]	23.60	0.808
JCD [33]	24.86	0.820
Ours	28.23	0.882

Table 1. Quantitative comparison against the state-of-the-art RSC methods on the proposed BS-RSC dataset.

Results on Fastec-RS. Besides the compassion on BS-RSC, we further evaluate the proposed method on the synthesized RSC dataset Fastec-RS to verify its effectiveness. The quantitative and qualitative results are shown in Tab. 2 and Fig. 7, respectively. We see that our model achieves comparable evaluation results against other methods.

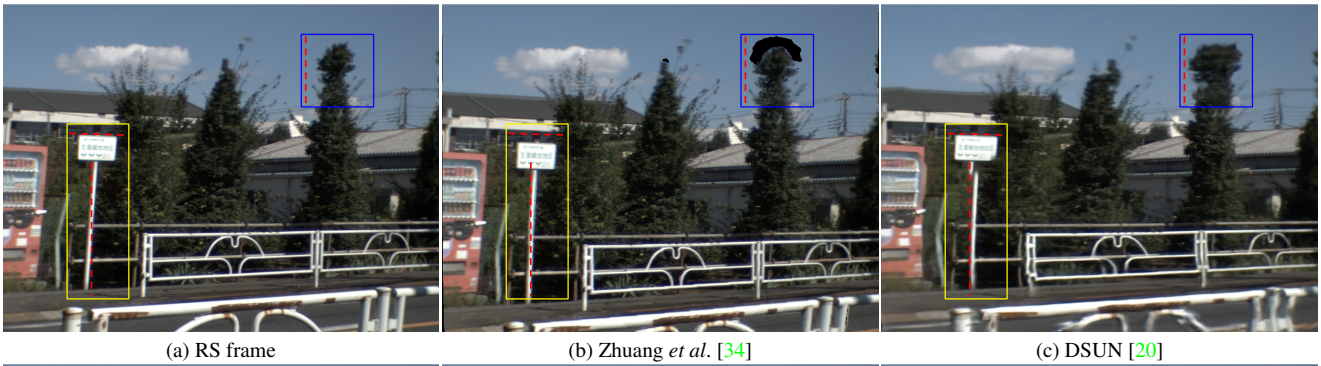
These quantitative and qualitative results shown above demonstrate the superior performance of our model.

Methods	PSNR↑(dB)	SSIM↑
Zhuang *et al*. [35]	21.44	0.710
DeepUnrollNet [20]	27.00	0.825
JCD [33]	24.84	0.778
SUNet [9]*	28.34	0.840
Ours	28.56	0.855

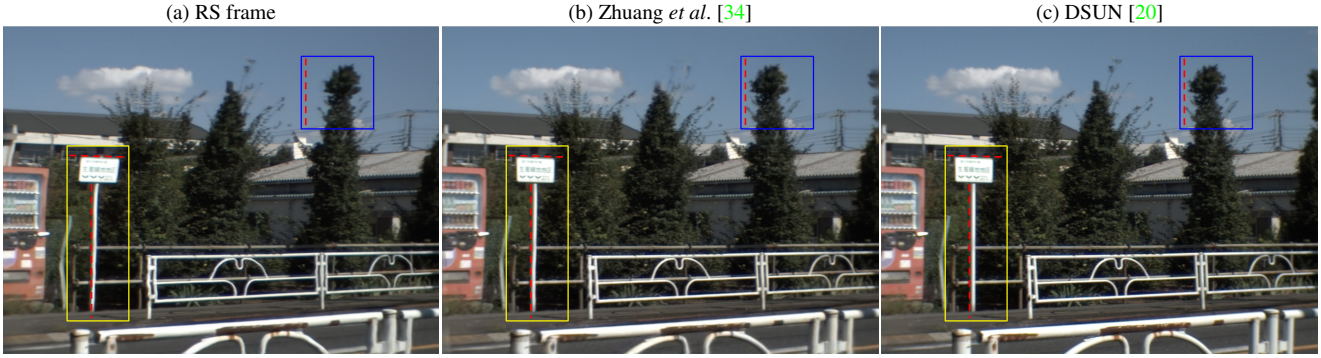
Table 2. Quantitative comparison against the state-of-the-art RSC methods on the synthesized Fastec-RC dataset. * means that SUNet restores the GS frame at the first row of the RS frame.

5.3. Ablation Study

Number of Input Frames. Our model takes three frames to model the motion information more accurately for warping. To verify this, we modify our model to adapt single frame and two frames input. Tab. 3 presents the quantitative results with different number of inputs. A single frame input achieves the lowest metrics due to the ill-posed nature

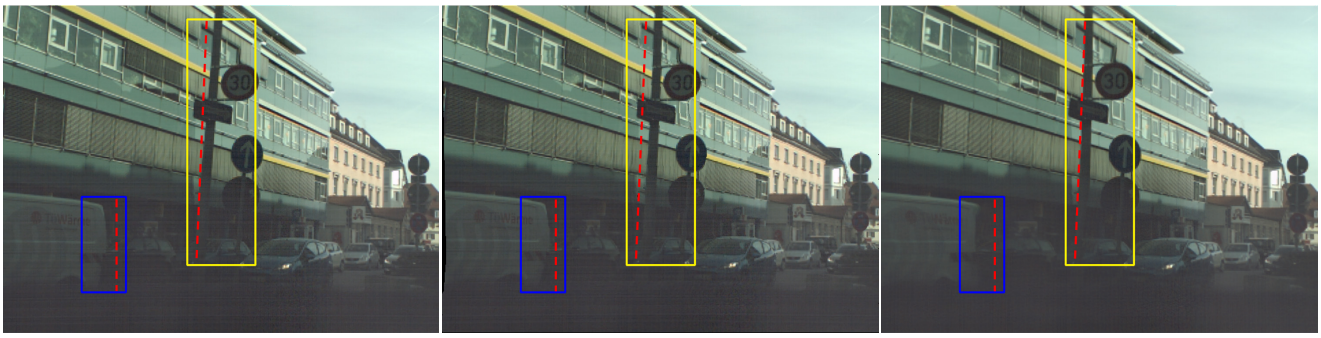


(a) RS frame (b) Zhuang *et al*. [34] (c) DSUN [20]

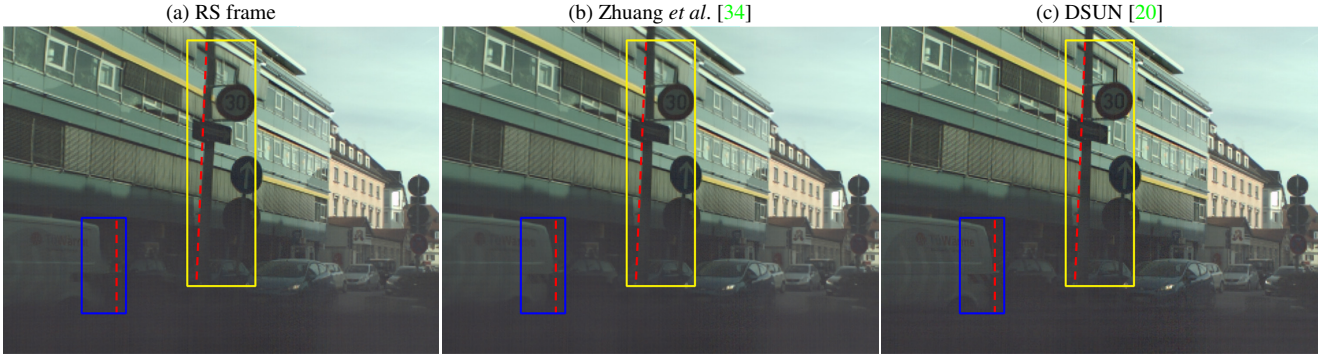


(d) JCD [33] (e) Ours (f) GS frame

Figure 6. Visual comparison on the proposed BS-RSC dataset for real-world RSC. Our method obtains higher visual quality, and more details are restored with fewer artifacts. Though the existing methods obtained highly competitive results on the synthesized dataset, they failed to restore the real-world RS distortions due to the difficulty of modeling the challenging motion in the BS-RSC.



(a) RS frame (b) Zhuang *et al*. [34] (c) DSUN [20]



(d) JCD [33] (e) Ours (f) GS frame

Figure 7. Visual results on the synthesized Fastec-RS dataset. The proposed method shows a strong competitive edge against other methods.

of estimating the displacement filed from a single frame. Instead, multi-frames can provide inter-frame movements and complementary information to perform better, especially when inputting the three consecutive RS frames.

Input Frames	PSNR↑(dB)	SSIM↑
1	23.84	0.765
2	27.20	0.838
3	28.56	0.855

Table 3. Ablation study of the number of the input RS frames.

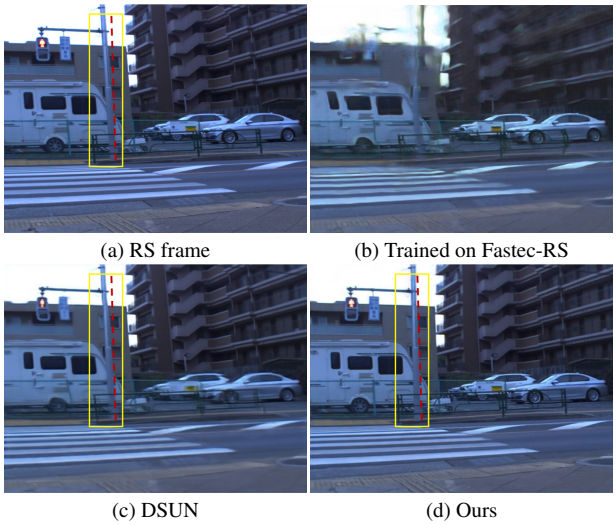
Adaptive Warping Module. To verify the effectiveness of the proposed warping module, we further construct three models. *Net1* only adopts a convolution for multiple RS features fusion without any warping. *Net2* replaces the AWM with the common backward warping. *Net3* replaces the AWM with the DFW adopted by existing methods. The results shown in Tab. 4 demonstrate the effectiveness of the proposed adaptive warping module. Meanwhile, our model achieves best PSNR and SSIM when number of the predicted motion fields M equals 9.

Model	PSNR↑(dB)	SSIM↑
Net1	26.14	0.801
Net2	26.76	0.826
Net3	27.20	0.837
Ours ($M = 2$)	27.41	0.836
Ours ($M = 9$)	28.56	0.855
Ours ($M = 16$)	27.98	0.850

Table 4. Ablation study of different warping methods.

Cross Camera Validation. To further validate the effectiveness of the proposed real-world RSC dataset BS-RSC, we test our model and DSUN model on the RS frames captured by third-party RS camera EO-1312C. The visual results are shown in Fig. 8. The sub-figure (b) losses many details compared to the original RS frame, which shows that the model trained on the synthesized dataset Fastec-RS cannot remove the RS effects and even introduces more blurs and artifacts into the image. Sub-figure (c) and (d) demonstrate that the proposed BS-RSC can help deal with real-world RS distortions. However, the DSUN model cannot estimate the displacement field effectively and correct the RS frame well. Thanks to the design of adaptive warping, our model obtains visually friendly results.

Inter-frame Time. To validate the generalisation capability of the proposed model, we test the trained model on the RS videos with different inter-frame time (by interpolating the RS frames), and the corrected results at different time stamps are shown in Fig. 9. We see that our model is robust



(a) RS frame (b) Trained on Fastec-RS

(c) DSUN (d) Ours

Figure 8. The corrected results of the proposed method on the frames captured by a third-party camera. **(a)** is the input RS frame. **(b)** is restored by our model trained on synthesized Fastec-RS. **(c)** and **(d)** are corrected by DSUN and our model trained on the proposed BS-RSC.

to different inter-frame time of the input RS frames during testing. However, some minor artifacts will occur (*e.g.*, the corrected GS frame with 1/4 inter-frame time) when the testing inter-frame time largely deviates from the inter-frame time of the training dataset. Therefore, we cannot perfectly avoid overfitting the inter-frame time of training dataset.

Figure 9. The corrected results of RS frames with different inter-frame time.

6. Limitation and Conclusion

In this paper, we explore the real-world rolling shutter correction task. An effective adaptive warping model based on the attention mechanism is proposed, and a real-world RSC dataset is collected by a well-designed beam-splitter acquisition system. Experimental results demonstrate the effectiveness of both, showing highly comparative results against previous warping-based methods. However, real-time inference on low-power mobile devices is still challenging at now, and how to accelerate the model is our future work.

Acknowledgments. This work was supported partially by the Major Research Plan of the National Natural Science Foundation of China (Grant No. 61991450), the Shenzhen Key Laboratory of Marine IntelliSense and Computation under Contract ZDSYS20200811142605016, and the JSPS KAKENHI with Grant Number 20H05951.

References

[1] Cenek Albl, Zuzana Kukelova, Viktor Larsson, and Tomas Pajdla. Rolling shutter camera absolute pose. *IEEE TPAMI*, 42(6):1439–1452, 2019. 1

[2] Cenek Albl, Zuzana Kukelova, Viktor Larsson, Michal Polic, Tomas Pajdla, and Konrad Schindler. From two rolling shutters to one global shutter. In *CVPR*, pages 2505–2513, 2020. 1, 3

[3] Cenek Albl, Zuzana Kukelova, and Tomas Pajdla. R6p-rolling shutter absolute camera pose. In *CVPR*, pages 2292–2300, 2015. 1

[4] Cenek Albl, Zuzana Kukelova, and Tomas Pajdla. Rolling shutter absolute pose problem with known vertical direction. In *CVPR*, pages 3355–3363, 2016. 1

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. 3

[6] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, pages 3086–3095, 2019. 5

[7] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, pages 12299–12310, 2021. 3

[8] Yuchao Dai, Hongdong Li, and Laurent Kneip. Rolling shutter camera relative pose: Generalized epipolar geometry. In *CVPR*, pages 4132–4140, 2016. 1

[9] Bin Fan, Yuchao Dai, and Mingyi He. Sunet: Symmetric undistortion network for rolling shutter correction. In *ICCV*, pages 4541–4550, 2021. 1, 2, 6

[10] Bin Fan, Ke Wang, Yuchao Dai, and Mingyi He. Rs-dpsnet: Deep plane sweep network for rolling shutter stereo images. *IEEE SPL*, 28:1550–1554, 2021. 1

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[13] Johan Hedborg, Per-Erik Forssén, Michael Felsberg, and Erik Ringaby. Rolling shutter bundle adjustment. In *CVPR*, pages 1434–1441, 2012. 1

[14] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019. 3

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[16] Yizhen Lao and Omar Ait Aider. Rolling shutter homography and its applications. *IEEE TPAMI*, 2020. 1

[17] Yizhen Lao and Omar Ait-Aider. A robust method for strong rolling shutter effects correction using lines with automatic feature selection. In *CVPR*, pages 4795–4803, 2018. 1

[18] Yizhen Lao, Omar Ait-Aider, and Adrien Bartoli. Rolling shutter pose and ego-motion estimation using shape-from-template. In *ECCV*, pages 466–482, 2018. 1

[19] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *CVPR Workshops*, pages 1833–1844, 2021. 3

[20] Peidong Liu, Zhaopeng Cui, Viktor Larsson, and Marc Pollefeys. Deep shutter unrolling network. In *CVPR*, pages 5941–5949, 2020. 1, 2, 3, 6, 7

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3

[22] Luc Oth, Paul Furgale, Laurent Kneip, and Roland Siegwart. Rolling shutter camera calibration. In *CVPR*, pages 1360–1367, 2013. 1

[23] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. In *NeurIPS*, 2019. 3

[24] Vijay Rengarajan, Yogesh Balaji, and AN Rajagopalan. Unrolling the shutter: Cnn to correct motion distortions. In *CVPR*, pages 2291–2299, 2017. 1, 2, 3

[25] Vijay Rengarajan, Ambasamudram N Rajagopalan, and Rangarajan Aravind. From bows to arrows: Rolling shutter rectification of urban scenes. In *CVPR*, pages 2773–2781, 2016. 1

[26] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, pages 184–201, 2020. 5

[27] Olivier Saurer, Kevin Koser, Jean-Yves Bouguet, and Marc Pollefeys. Rolling shutter stereo. In *CVPR*, pages 465–472, 2013. 1

[28] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018. 1, 2, 4

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 3

[30] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 3

[31] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, pages 5791–5800, 2020. 3

[32] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *ECCV*, pages 191–207, 2020. 5

[33] Zhihang Zhong, Yinqiang Zheng, and Imari Sato. Towards rolling shutter correction and deblurring in dynamic scenes. In *CVPR*, pages 9219–9228, 2021. 2, 5, 6, 7

[34] Bingbing Zhuang, Loong-Fah Cheong, and Gim Hee Lee. Rolling-shutter-aware differential sfm and image rectification. In *CVPR*, pages 948–956, 2017. 1, 6, 7

[35] Bingbing Zhuang, Quoc-Huy Tran, Pan Ji, Loong-Fah Cheong, and Manmohan Chandraker. Learning structure-and-motion-aware rolling shutter correction. In *CVPR*, pages 4551–4560, 2019. 1, 2, 3, 6