Towards Rolling Shutter Correction and Deblurring in Dynamic Scenes

Zhihang Zhong[1,2]   Yinqiang Zheng[1]   Imari Sato[2,1]

[1]The University of Tokyo, Japan   [2]National Institute of Informatics, Japan
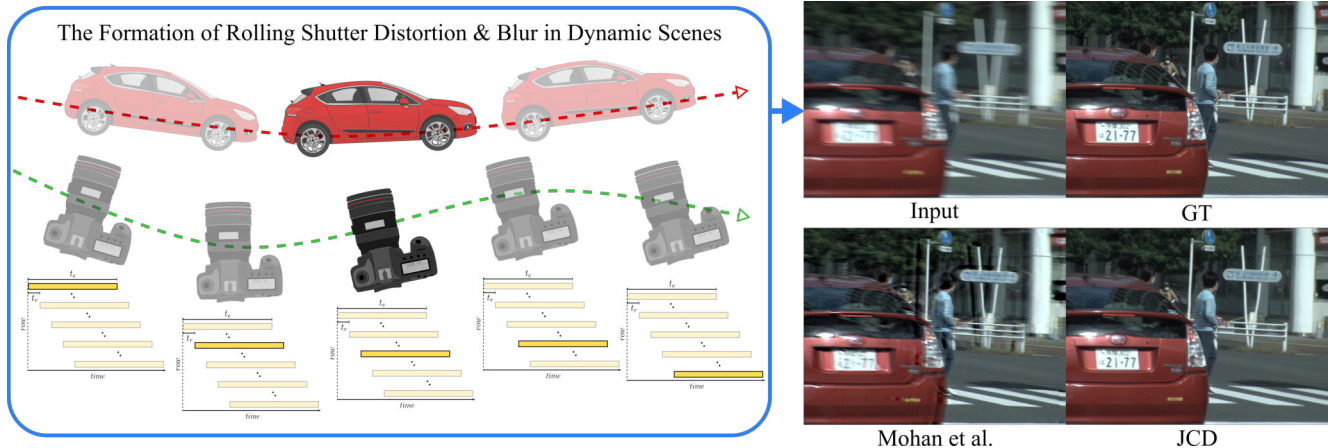
Figure 1: **Joint rolling shutter correction and deblurring (RSCD)**. Image degradation caused by joint rolling shutter distortion and motion blur becomes extremely complex, when both ego-motion and object-motion are involved in dynamic scenes. Existing methods for RSCD such as Mohan *et al.* [22] are developed for static scenes. We present the first realistic benchmark dataset (BS-RSCD) and model (JCD) to advance learning-based approaches in this field.

Abstract

Joint rolling shutter correction and deblurring (RSCD) techniques are critical for the prevalent CMOS cameras. However, current approaches are still based on conventional energy optimization and are developed for static scenes. To enable learning-based approaches to address real-world RSCD problem, we contribute the first dataset, BS-RSCD, which includes both ego-motion and object-motion in dynamic scenes. Real distorted and blurry videos with corresponding ground truth are recorded simultaneously via a beam-splitter-based acquisition system.

Since direct application of existing individual rolling shutter correction (RSC) or global shutter deblurring (GSD) methods on RSCD leads to undesirable results due to inherent flaws in the network architecture, we further present the first learning-based model (JCD) for RSCD. The key idea is that we adopt bi-directional warping streams for displacement compensation, while also preserving the non-warped deblurring stream for details restoration. The experimental results demonstrate that JCD achieves state-of-the-art performance on the realistic RSCD dataset (BS-RSCD) and the synthetic RSC dataset (Fastec-RS). The

dataset and code are available at `https://github.com/zzh-tech/RSCD`*.*

1. Introduction

Rolling shutter (RS) CMOS cameras dominate the consumer market, especially the mobile phone market, due to their largely reduced power consumption, low cost and compact design [13]. However, if the imaging conditions are not ideal, video recorded through RS mechanism will suffer from compound quality issues. Specifically, the RS distortion (Jello effect) and motion blur become noticeable when there is a large relative motion between the camera and the object, especially in poorly illuminated environment where longer exposure time is required. Joint rolling shutter correction and deblurring (RSCD) techniques are seldom explored and are urgently needed for RS-based devices.

Existing works generally treat rolling shutter correction (RSC) and deblurring as two separate issues. As for RSC methods [30, 29, 47, 18, 37, 20], they assume by default

---

that there are no blur effects in the captured image but only distortion caused by the scanning strategy of RS that exposes each row sequentially. The formation of the image I_r with only RS distortion can be described as follows:

$$I_r^{(t)}[i] = I_g^{(t-t_m+it_r)}[i], \qquad (1)$$

where t_m equals $(M/2)t_r$; t_r denotes the readout (offset) time for each row of RS; M denotes total number of rows of the image; $I_r^{(t)}[i]$ denotes the ith row of the RS distorted image I_r with the middle moment of exposure at time t; $I_g^{(t-t_m+it_r)}[i]$ denotes the same row of the virtual global shutter (GS) image I_g with the middle moment of exposure at time $t-t_m+it_r$, the entire scene of which is captured simultaneously. On the other hand, existing deblurring methods (GSD) [40, 10, 33, 24, 35, 25, 26, 44] typically assume the target image is captured by GS. Then, the formation of the GS blurry image I_b is given by:

$$I_b^{(t)} = \frac{1}{t_e} \int_{t-t_e/2}^{t+t_e/2} I_g^{(t)} dt, \qquad (2)$$

where t_e denotes the exposure time of GS. The combined effect of RS distortion and blur escalates the problem to a new dimension. The formation of RS distorted and blurry image I_{rb} can be described as follows:

$$I_{rb}^{(t)}[i] = \frac{1}{t_e} \int_{t-t_m+it_r-t_e/2}^{t-t_m+it_r+t_e/2} I_g^{(t-t_m+it_r)}[i] dt. \qquad (3)$$

As the reverse process of the Eq. (3), RSCD is extremely challenging because it requires estimating pixel-level displacements and blur kernels simultaneously.

Under simplified or idealized conditions, there are few works [34, 22] that specifically address blind RSCD problem. Assuming a static scene and negligible in-plane rotation, Su *et al.* [34] propose RS-BMD to deliver distortion-free and sharp image by estimating parametric trajectory of the camera. Mohan *et al.* [22] further remove the limitation of parametric trajectory estimation, allowing their method to handle RS blurry image produced by irregular camera trajectory. However, these methods cannot cope with freely moving objects in dynamic scenes, which are often observed in a real-world scenario. Due to the complexity of relative motion in the RSCD problem, not only is RSCD modeling difficult, but inferring the latent image is also time-consuming.

Recently, the success of deep learning methods and the corresponding large-scale datasets has greatly facilitated the development of image and video restoration techniques. However, the development of learning-based RSCD methods is still hampered by the lack of datasets. Even for pure RSC problem, there is only one public dataset (Fastec-RS [20]) available for data-driven methods, which synthesizes RS images by sequentially copying a row of pixels from captured high-FPS GS images. It still remains a challenge to obtain RS distorted and blurry video and the corresponding GS sharp video for the same scenes in the wild.

In this paper, to enable data-driven methods for RSCD, we propose BS-RSCD, the first dataset used for real-world RSCD task, using a well-designed beam-splitter acquisition system. A RS camera and a GS camera are physically aligned to capture RS distorted and blurry as well as GS sharp video pairs simultaneously. Based on this dataset, we further explore the possibility of applying deep learning to address realistic RSCD problem. The complexity of BS-RSCD brings a new challenge to the existing neural network architecture of GSD or RSC. Through experiments, we found that the existing GSD methods are prone to destroy the original geometric structure of the scene when facing the displacement introduced by RS distortion, while the existing RSC methods are difficult to recover the details from motion blur. To solve this dilemma, we design a novel joint correction and deblurring model (JCD) that incorporates the advantages of both RSC and GSD neural network architectures. To achieve a better trade-off between distortion correction and motion deblurring, JCD fuses bidirectional warped features and non-warped deblurring features at multiple scales using a deformable attention module. Our contributions can be summarized as follows:

- We introduce BS-RSCD, the first dataset for joint RSCD problem in real dynamic scenes, using a beam-splitter acquisition system.

- We propose a novel neural network architecture that can handle both RS distortion and blur, using deformable attention module to fuse features from bidirectional warping and deblurring streams.

- Experimental results demonstrate the superiority of the proposed method over the state-of-the-art methods in both RSC and RSCD tasks, and show the effectiveness of our BS-RSCD dataset.

2. Related works

2.1. Global Shutter Deblurring

Early works such as Fergus *et al.* [3] only consider the uniform blur caused by camera shake. The camera motion is estimated for uniform blur kernel inference. Considering blur of real scenes is spatially varying, Levin *et al.* [19] and Wulff *et al.* [40] segment the image into different regions for distinct deconvolution operation with segmentation-wise blur kernel. Then, Hyun *et al.* [9, 10] use energy model to estimate pixel-wise blur kernel without segmentation. Ren *et al.* [28] introduce non-linear optical flow for more accurate pixel-wise blur kernel estimation.

Recently, data-driven approaches achieve leading results for image and video deblurring with an end-to-end manner.

Deep learning models are widely used to directly predict sharp image without blur kernel estimation. Nah *et al.* [24] propose a CNN model to remove blur following a coarse-to-fine pyramid architecture. Tao *et al.* [35] further improve the architecture by proposing scale-recurrent structure with ConvLSTM [32]. In STRCNN [33] and EDVR [38], multiple neighboring images of the blurry video are used to provide temporal correlation for deblurring. To better utilize the temporal information, DBN [11], STFAN [45], IFIRNN [25], ESTRNN [44] and CDVD-TSP [26] adopt the recurrent structure from RNN and achieve impressive deblurring performance. In addition, Kupyn *et al.* [16, 17] adopt generative adversarial networks to produce smoother and more visually appealing results.

However, the aforementioned image and video deblurring methods have a common implicit assumption that the camera uses a global shutter. These methods will fail when large and complicated displacements caused by the RS distortion appear in the blurry image.

2.2. Rolling Shutter Correction

To solve RSC problem, Forssen *et al.* [5] model the camera motion as a parametrised continuous curve and solve parameters using non-linear least squares over inter-frame correspondences. Grundmann *et al.* [6], Liu *et al.* [21] and Lao *et al.* [18] address the RSC problem with the help of RANSAC [4]. Rengarajan *et al.* [30] correct the RS distorted image with the aid of straight line assumption. Zhuang *et al.* [47] estimate the depth map and motion for RSC by solving a SfM problem using two consecutive RS images. Vasu *et al.* [37] propose a new pipeline to sequentially recover both motion of fast moving camera and scene structure from depth-dependent RS distortion. Recently, Albl *et al.* [1] present an simple yet effective method to restore a GS image by using dual RS images with opposite distortion of a same scene.

Inspired by the success of deep learning models, several deep neural network-based methods [29, 48, 20] are proposed and show impressive performance. Rengarajan *et al.* [29] propose a CNN model with long rectangular convolutional kernel to estimate the row-wise camera motion for undoing RS distortion. Zhuang *et al.* [48] follow the scheme of [47] but use two independent neural network branches to estimate dense depth map and camera motion, respectively. Liu *et al.* first propose an end-to-end network for RSC using a differentiable forward warping module.

Current solutions of the RSC methods cannot handle the blur well presented in RS distorted images even if they are learning-based.

2.3. Joint Correction and Deblurring

Few works consider the challenging situation that joint RS distortion and blur appear in the images simultane-ously. Tourani *et al.* [36] use feature matches between depth maps to timestamp parametric ego-motion to further achieve RSCD. Their method needs multiple RGBD images as inputs. Hu *et al.* [8] use information of inertial sensor in smartphone instead to estimate ego-motion of the RGB images with RS distortion and blur. Su *et al.* [34] first propose an energy model, RS-BMD, to directly estimate the ego-motion without device-specific constraints. However, they discard in-plane rotation to simplify ego-motion. Mohan *et al.* [22] overcome the inability of 3D rotation and irregular ego-motion of RS-BMD. They leverage a generative model for removing RS motion blur and a prior to disambiguate multiple solutions during inversion.

Existing RSCD methods are designed for static scenes with only ego-motion, and there are no learning-based methods for RSCD problem due to the lack of datasets.

3. BS-RSCD Dataset

It is worth noting that the current mainstream GSD and RSC datasets are synthetic. As for popular GSD datasets, GOPRO [24], DVD [33] and REDS [23] accumulate consecutive sharp images from high-FPS GS video to generate a blurry image. As for RSC dataset, Fastec-RS [20] uses a GS camera mounted on a ground vehicle to capture high-FPS video with only horizontal motion. Then, RS images are synthesized by sequentially copying a row of pixels from consecutive GS images. Deep learning models trained on the synthetic dataset have limited performance for the data in realistic environment, because the distribution of synthetic data differs from the real data and there may be artifacts. In the case of Fastec-RS, the exposure intensity of the pixel rows from different GS images is inconsistent, resulting in obvious horizontal striping artifacts in the synthetic input, as illustrated in the 1st row of Fig. 3.

To improve the generality of deep learning models, real datasets without synthetic artifacts are essential. Recently, researchers have succeeded in designing specific optical acquisition system to capture real pairs of images or videos for training neural networks. Zhang *et al.* [42] use a zoom lens to collect low-resolution and super-resolution image

	Train	Validation	Test
Sequences	50	15	15
Frames/seq.	50	50	50
Frames	2500	750	750
Resolution		640×480	
RS camera		FLIR FL3-U3-13S2C	
GS camera		FLIR GS3-U3-28S4C	

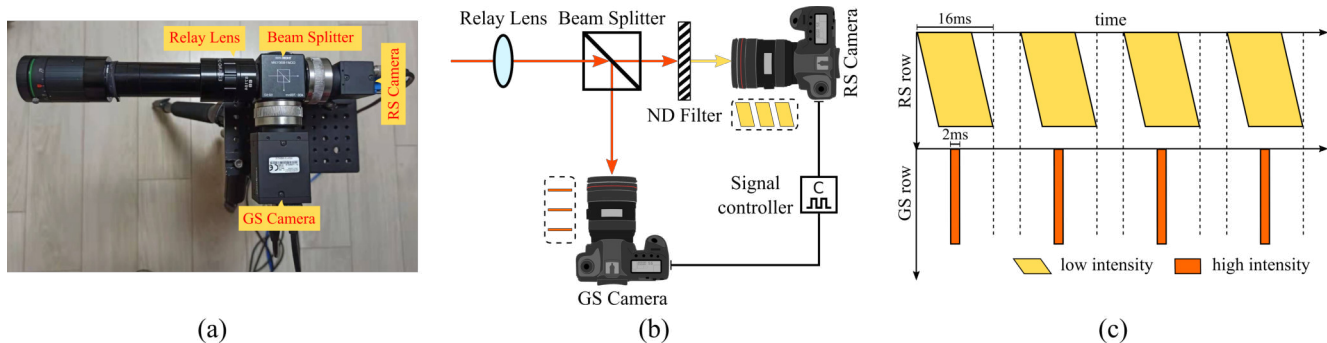Table 1: **Configuration of the proposed BS-RSCD.**

Figure 2: **Beam-splitter acquisition system.** (a) shows real system used to collect the dataset; (b) is system schematic diagram; (c) is exposure scheme of the system.
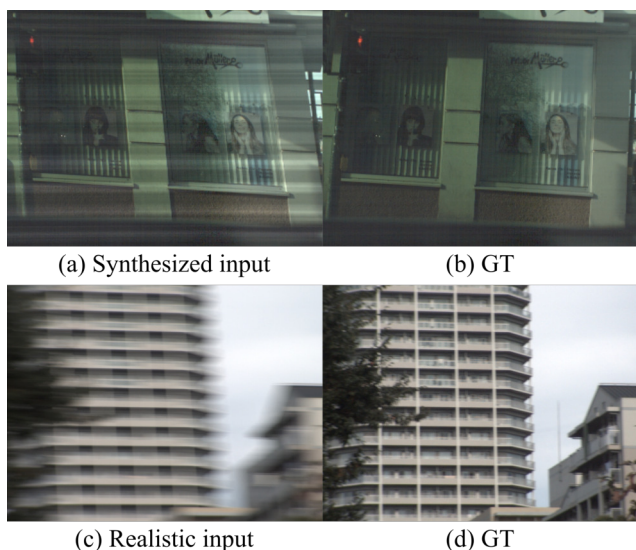

Figure 3: **Data samples of Fastec-RS and BS-RSCD.** 1st row shows unrealistic case of Fastec-RS with striping artifacts. 2nd row shows natural RS distortion and blur in BS-RSCD.

pairs for static scenes. Rim *et al.* [31] and Zhong *et al.* [44] adopt beam-splitter acquisition system to collect single image and video deblurring datasets, respectively. Inspired by the above works, we also propose a beam-splitter acquisition system (Fig. 2) to collect the first dataset for RSCD problem, named BS-RSCD. Data samples from the captured video pairs are shown in the 2nd row of Fig. 3. The RS distortion and motion blur in the realistic input look more natural.

3.1. Hardware Configuration

As for the hardware, we choose a RS camera FLIR FL3-U3-13S2C equipped with a 1/3-inch CMOS sensor, whose pitch size is $3.63\,\mu m$. To properly match the sensor specifications of the RS camera, a GS camera FLIR GS3-U3-

28S4C is used, whose sensor size is 1/1.8-inch, and pitch size is $3.69\,\mu m$. These two cameras are co-axis aligned via a 50/50 beam splitter. As for geometric alignment, the two cameras are first mechanically aligned assisted with collimated laser beams. Later, a homography correction using standard checker pattern is conducted, so as to reduce the alignment error to less than one pixel. The exposure time of the RS and the GS camera is 16ms and 2ms, respectively. Both cameras run at 15fps. We use a wave generator to generate synchronized pulses at 15Hz, and the phase of the pulse for the GS camera is properly delayed, such that the GS exposure time lies in the middle of the RS exposure duration (Fig. 2(c)). As for photometric alignment, we put a neutral density filter before the RS camera, such that the intensity values of two cameras are almost equalized. We further use a color checker pattern to correct the RGB response of the GS camera, such that both cameras record the same RGB images, when the scene and cameras are static.

3.2. Collected Data

We collected BS-RSCD in the dynamic urban environment with both ego-motion and object-motion. The configuration of BS-RSCD is listed in Table 1. There are 80 short video sequences of 50 frames each in our dataset (4000 images pairs in total). We divided the BS-RSCD into 50, 15 and 15 sequences representing the training set (2500 image pairs), the validation set (750 image pairs) and the test set (750 image pairs), respectively. The original resolution (1280×920) was downsampled to half (640×460) to suppress the noise level of ground truth.

4. Method

Through experiments on BS-RSCD, we found that existing GSD and RSC methods suffer from inherent flaws caused by the design of their neural network architecture, leading to undesirable geometric distortions or blurry details. As for the end-to-end RSC method [20], latent undistorted image is inferred based on warped features. Learn-
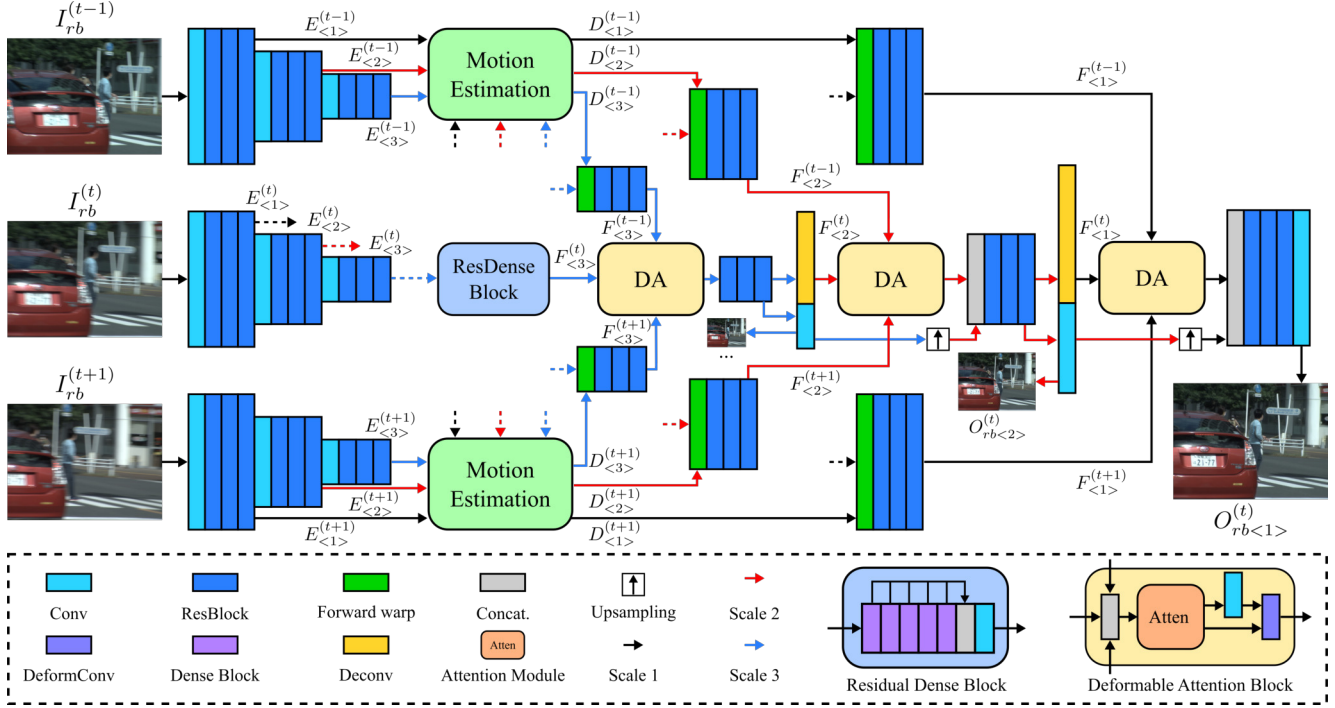
Figure 4: **Neural network architecture of JCD.** It consists of three streams, including forward and backward warping streams for distortion compensation, and a deblurring stream for detail restoration. A deformable attention block is applied to the concatenation of warped and deblurring features at each scale to predict images in a coarse-to-fine manner.

ing how to deblur only from warped features is difficult because some of the original information is lost and the process of warping may be imprecise. On the other hand, the encoder-decoder network architecture of the GSD methods fails to preserve the geometry of the objects in the presence of RS distortion. Thus, we propose a novel network for joint correction and deblurring, named JCD, that combines the advantages of RSC and GSD neural network architectures to effectively compensate the displacement and restore the details, shown in Fig. 4. The entire structure consists of three streams taking current frame I_{rb}^t and two nearby frames $\left\{ I_{rb}^{(t-1)}, I_{rb}^{(t+1)} \right\}$ as inputs. The encoder features of the nearby frames are used to produce the warped features of the current frame. In the decoder stage, the features from three streams are fused to predict latent image $O_{rb<1>}^t$ from coarse to fine.

4.1. Bi-directional Warping Streams and Middle Deblurring Stream

First, to promote the precision of warping process, we use bi-directional warping streams to warp different scales of encoder features $\left\{ E_{<3>}^{(t)}, E_{<2>}^{(t)}, E_{<1>}^{(t)} \right\}$ of current frame from two directions, forward and backward. We adopt the motion estimation module and differentiable forward warping (DFW) block of [20] to realize both forward

warping and backward warping processes. DFW warps features of current frame based on the displacement field ($D_{<s>}^{(t-1)}$ or $D_{<s>}^{(t+1)}$) that is estimated by the correlation layer [12] between the same scale features of current frame ($E_{<s>}^{(t)}$) and nearby frame ($E_{<s>}^{(t-1)}$ or $E_{<s>}^{(t+1)}$). The forward warped features $F_{<s>}^{(t-1)}$ of each scale are given by:

$$F_{<s>}^{(t-1)} = DFW\left(E_{<s>}^{(t)}, D_{<s>}^{(t-1)} \right), s \in \{1, 2, 3\}. \quad (4)$$

Similarly, the backward warped features are given by:

$$F_{<s>}^{(t+1)} = DFW\left(E_{<s>}^{(t)}, D_{<s>}^{(t+1)} \right), s \in \{1, 2, 3\}. \quad (5)$$

Another advantage of bi-directional warping is that it avoids the need for the model to guess and generate the information of occluded regions around the image borders. Instead, it incorporates the corresponding information into either the forward or backward warped features. In addition to the bi-directional warping streams, we also add an encoder-decoder stream commonly used in deblurring neural networks, denoted as middle deblurring stream. The warping-free features of the middle deblurring stream are more conducive for the model to learn the sharp details of the latent image. The deblurring stream is connected through a residual dense block [43], which further extracts rich spatial features $F_{<3>}^{(t)}$ for current frame with different receptive fields.

4.2. Deformable Attention Module

In order to effectively integrate the features of bi-directional warping and deblurring streams, we propose a fusion module as deformable attention (DA) block. The DA block handles the concatenation of features $(CAT(F_{<s>}^{(t-1)}, F_{<s>}^{(t)}, F_{<s>}^{(t+1)}))$ from different streams via an attention module followed by a deformable convolutional layer [2, 46]. Channel attention module (squeeze-and-excitation block [7]) is adopted to assign channel-wise weight to the concatenated features based on their importance. Since features from warping streams and deblurring stream are not spatially consistent, it is difficult to leverage the information from weighted features using normal 2D convolutional layers. Instead, we use deformable convolution with learnable offset to further fuse the features, which is able to dynamically adjust its receptive field to handle geometric transformation.

4.3. Loss Function

The inference process uses a pyramid structure in a coarse-to-fine manner. There is a predicted result for each scale as $\{O_{rb<3>}^{(t)}, O_{rb<2>}^{(t)}, O_{rb<1>}^{(t)}\}$. The Charbonnier loss and perceptual loss [14] are used to constrain the predicted results at all scales, while total variation loss is applied to the estimated displacement fields to smooth the forward and backward warping processes. The total loss function is given by:

$$\mathcal{L}_{total} = \sum_{s=1}^{3} (\lambda_c \mathcal{L}_c + \lambda_p \mathcal{L}_p + \lambda_{tv} \mathcal{L}_{tv}). \qquad (6)$$

5. Experiments and Results

5.1. Implementation Details

We implemented our model with PyTorch [27]. The network is trained in 400 epochs with batch size and learning rate as 8 and 3×10^{-4}. λ_c, λ_p and λ_{tv} are set as 10, 1, 0.1, respectively. Random cropping (256×256) is applied for data augmentation. ADAM [15] is used to update weight with cosine annealing scheduler. We adopt the deformable convolution layer from Detectron2 [39] with deformable groups as 8. Standard metrics PSNR and SSIM, as well as learned perceptual metric LPIPS [41] are used for quantitative evaluation.

5.2. Results

5.2.1 Results on BS-RSCD

We compare JCD against state-of-the-art (SoTA) RSC [20], GSD [25, 44] and RSCD [34, 22] methods on BS-RSCD. The qualitative results are presented in Fig. 5. The predicted results of GSD method, ESTRNN [44], cannot preserve the

	PSNR ↑	SSIM ↑	LPIPS ↓
IFIRNN [25]	25.77	0.753	0.312
ESTRNN [44]	26.90	0.784	0.264
DeepUnrollNet [20]	25.14	0.729	0.159
RS-BMD [34]	15.18	0.316	0.543
Mohan *et al.* [22]	17.46	0.420	0.404
JCD	26.42	0.757	0.122

Table 2: **Quantitative results on BS-RSCD dataset.**

	PSNR ↑	SSIM ↑	LPIPS ↓
ESTRNN [44]	27.41	0.844	0.189
Zhuang *et al.* [47]	21.44	0.710	0.218
DeepUnrollNet [20]	27.00	0.825	0.108
JCD	24.84	0.778	0.107

Table 3: **Quantitative results on Fastec-RS dataset.**

geometric structure of the object resulting in distorted artifacts. RSC method [20] fails to recover the details of some blurry areas. The performance of SoTA RSCD method is even worse than the learning-based GSD and RSC methods due to the limited modeling capabilities of conventional energy optimization and the assumption of static scenes. Our method achieves the most visually pleasing results, with details and geometry close to the ground truth.

Quantitative comparison is presented in Table 2. We found that the learned perceptual metrics LPIPS can better reflect the quality of results than SSIM and PSNR for tasks related to RS distortion correction, because LPIPS is less sensitive to pixel position. GSD methods [25, 44] without warping process are prone to forcing only some of pixels to the correct position and value, thus cheating non-perceptual metrics (PSNR and SSIM) and obtaining higher values. However, their visual results are much less visually appealing and the LPIPS performance is poor. For example, ESTRNN achieves much higher PSNR (26.90 dB) and SSIM (0.784) than other methods but only 0.264 for LPIPS (lower LPIPS is better). Our method achieves best LPIPS performance as 0.122.

5.2.2 Results on Fastec-RS

In addition, we also compare JCD with the SoTA RSC methods [47, 20] on synthetic RSC dataset, Fastec-RS. The qualitative and quantitative results are presented in Fig. 6 and Table 3, respectively. Compared to DeepUnrollNet [20], our method achieves better visual performance due to more accurate displacement estimation from the bi-
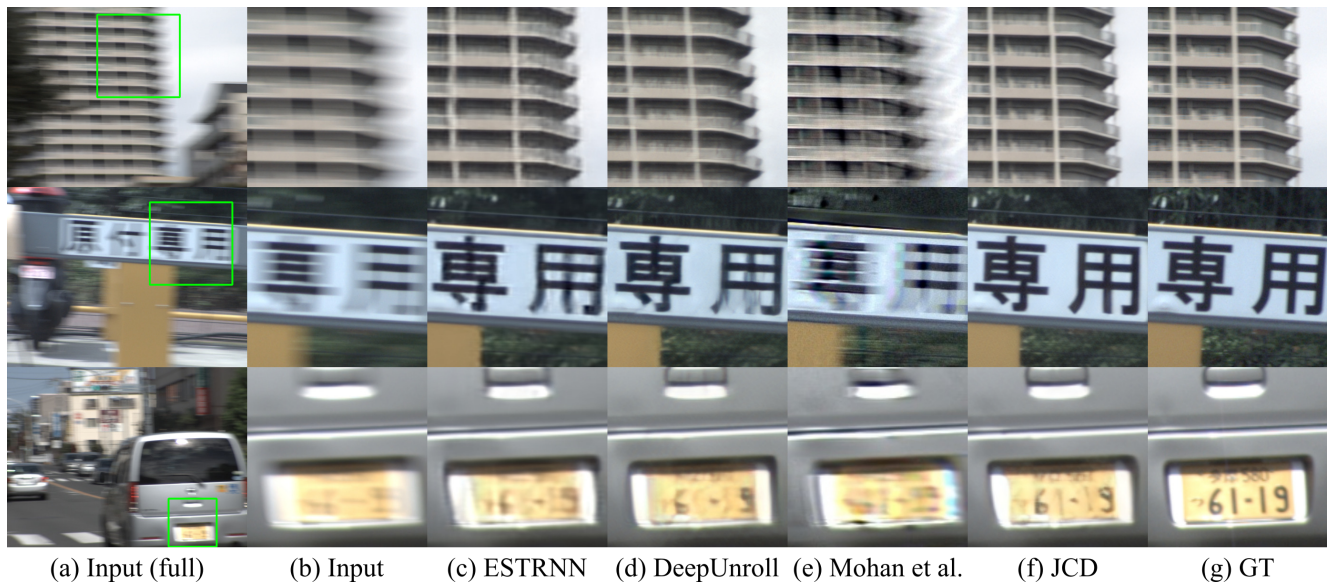
(a) Input (full)   (b) Input   (c) ESTRNN   (d) DeepUnroll (e) Mohan et al.   (f) JCD   (g) GT

Figure 5: **Visual results on BS-RSCD dataset.**

(a) Input (full)   (b) Input   (c) ESTRNN (d) Zhuang et al. (e) DeepUnroll   (f) JCD   (g) GT
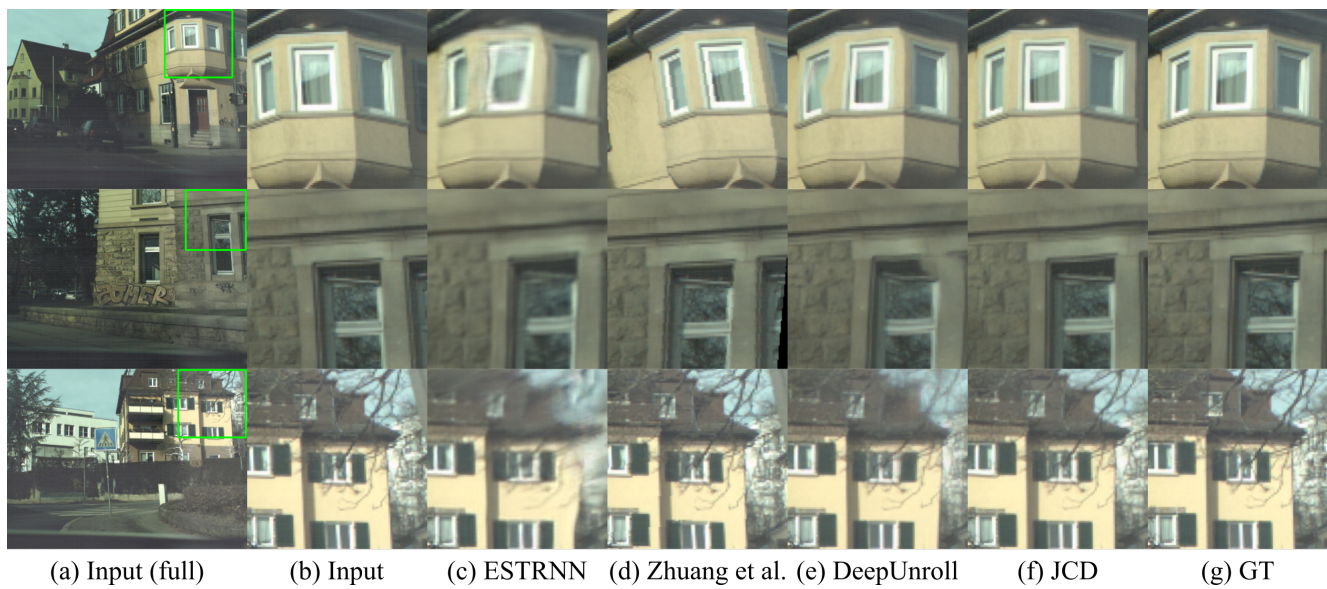
Figure 6: **Visual results on Fastec-RS dataset.**

	BS-RSCD			Fastec-RS		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR↑	SSIM ↑	LPIPS ↓
JCD (w/o ms)	25.02	0.721	0.181	25.52	0.784	0.126
JCD (w/o dcn)	25.51	0.736	0.134	25.32	0.785	0.115
JCD (w/o bs)	25.47	0.736	0.127	24.36	0.769	0.112
JCD	26.42	0.757	0.122	24.84	0.778	0.107

Table 4: **Quantitative ablation study on BS-RSCD and Fastec-RS.**

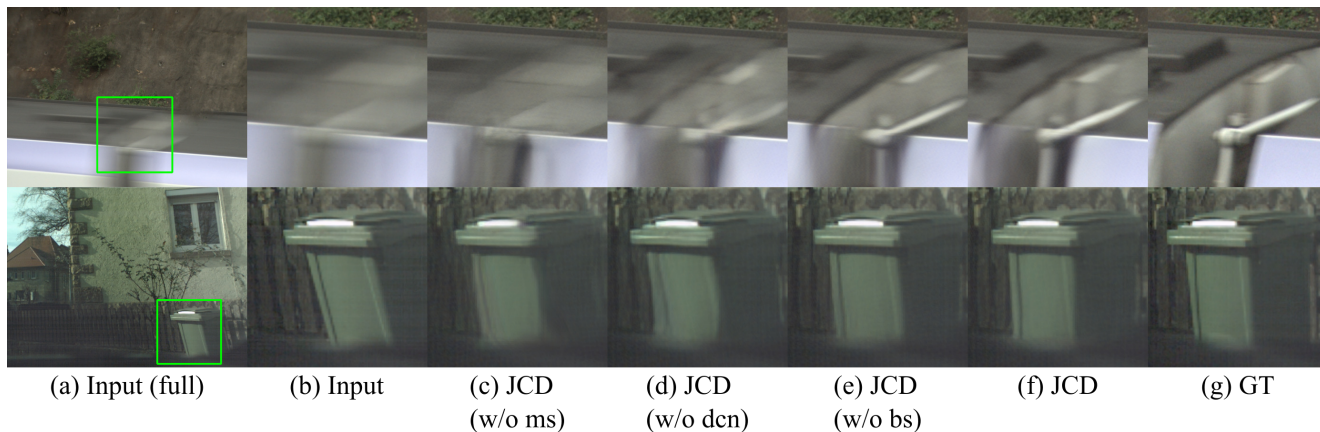|(a) Input (full)|(b) Input|(c) JCD (w/o ms)|(d) JCD (w/o dcn)|(e) JCD (w/o bs)|(f) JCD|(g) GT|

Figure 7: **Visual results of ablation study.** 1st row shows results on BS-RSCD; while 2nd row shows results on Fastec-RS.

(a) Input   (b) JCD

Figure 8: **Test on the data of third-party cameras.** 1st row shows the results of FLIR camera; 2nd row shows the results of EO camera.

directional warping streams and sharper details from the middle deblurring stream. We also test GSD method (ES-TRNN [44]) on Fastec-RS and observed the same phenomena that the results of ESTRNN have much higher PSNR and SSIM scores but worse visual results and LPIPS score. Compared with DeepUnrollNet, the proposed method has slightly improved in terms of LPIPS, while the improvement in visual performance is more obvious. The visual results of DeepUnrollNet as illustrated in Fig. 6(e) exist some undesired distortion and blur.

5.3. Ablation Study

We also implemented ablation study experiments to validate the effectiveness of proposed neural network architecture. We implemented other three models including no the middle deblurring stream (w/o ms), no deformable convolution layers (w/o dcn), and only one warping stream (w/o bs). The qualitative and quantitative results on both BS-RSCD and Fastec-RS are presented as Fig. 7 and Table 4. The results demonstrate that deblurring stream is important for recovering details, the deformable convolutional layers in the fusion module reduce the undesired variations in object geometry, and the bi-directional warping streams further improve the compensation for displacements and details.

5.4. Camera Cross Validation

Finally, we tested the data from two third-party RS cameras, EO 1312LE and FLIR BFLY-PGE-14S2C, using the trained model on BS-RSCD. The visual results are presented in Fig. 8. It demonstrates that model trained on our BS-RSCD is generic cross different devices.

6. Conclusion

In this paper, we constructed the first dataset (BS-RSCD) for joint rolling shutter correction and deblurring task, using a beam-splitter acquisition system. We also proposed the first learning-based method JCD that incorporates the advantages of the network architectures from pure rolling shutter correction and global shutter deblurring tasks. Experimental results demonstrated that fusing the features from both bi-directional warping streams and deblurring stream, using deformable attention module, can restore sharp details and preserve geometric structure of objects in the dynamic scenes. Collecting a larger dataset and proposing a more reasonable and interpretable metrics for the RSCD and RSC tasks will be our future works.

Acknowledgement

References

[1] Cenek Albl, Zuzana Kukelova, Viktor Larsson, Michal Polic, Tomas Pajdla, and Konrad Schindler. From two rolling shutters to one global shutter. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2505–2513, 2020.

[2] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[3] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis, and William T Freeman. Removing camera shake from a single photograph. In *ACM SIGGRAPH 2006 Papers*, pages 787–794. 2006.

[4] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

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

[6] Matthias Grundmann, Vivek Kwatra, Daniel Castro, and Irfan Essa. Calibration-free rolling shutter removal. In *2012 IEEE international conference on computational photography (ICCP)*, pages 1–8. IEEE, 2012.

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[8] Zhe Hu, Lu Yuan, Stephen Lin, and Ming-Hsuan Yang. Image deblurring using smartphone inertial sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1855–1864, 2016.

[9] Tae Hyun Kim and Kyoung Mu Lee. Segmentation-free dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2766–2773, 2014.

[10] Tae Hyun Kim and Kyoung Mu Lee. Generalized video deblurring for dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5426–5434, 2015.

[11] Tae Hyun Kim, Kyoung Mu Lee, Bernhard Scholkopf, and Michael Hirsch. Online video deblurring via dynamic temporal blending network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4038–4047, 2017.

[12] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[13] James Janesick, Jeff Pinter, Robert Potter, Tom Elliott, James Andrews, John Tower, John Cheng, and Jeanne Bishop. Fundamental performance differences between cmos and ccd imagers: part iii. In *Astronomical and Space Optical Systems*, volume 7439, page 743907. International Society for Optics and Photonics, 2009.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[17] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[18] Yizhen Lao and Omar Ait-Aider. A robust method for strong rolling shutter effects correction using lines with automatic feature selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4795–4803, 2018.

[19] Anat Levin. Blind motion deblurring using image statistics. In *Advances in Neural Information Processing Systems*, pages 841–848, 2007.

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

[21] Shuaicheng Liu, Lu Yuan, Ping Tan, and Jian Sun. Bundled camera paths for video stabilization. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013.

[22] Mahesh MR Mohan, AN Rajagopalan, and Gunasekaran Seetharaman. Going unconstrained with rolling shutter deblurring. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4010–4018, 2017.

[23] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[24] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017.

[25] Seungjun Nah, Sanghyun Son, and Kyoung Mu Lee. Recurrent neural networks with intra-frame iterations for video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8102–8111, 2019.

[26] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3043–3051, 2020.

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming

Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[28] Wenqi Ren, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Video deblurring via semantic segmentation and pixel-wise non-linear kernel. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1077–1085, 2017.

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

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

[31] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[32] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28:802–810, 2015.

[33] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1279–1288, 2017.

[34] Shuochen Su and Wolfgang Heidrich. Rolling shutter motion deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1529–1537, 2015.

[35] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018.

[36] Siddharth Tourani, Sudhanshu Mittal, Akhil Nagariya, Visesh Chari, and Madhava Krishna. Rolling shutter and motion blur removal for depth cameras. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 5098–5105. IEEE, 2016.

[37] 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*, pages 636–645, 2018.

[38] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[39] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[40] Jonas Wulff and Michael Julian Black. Modeling blurred video with layers. In *European Conference on Computer Vision*, pages 236–252. Springer, 2014.

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[42] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2019.

[43] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.

[44] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020.

[45] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2482–2491, 2019.

[46] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019.

[47] 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*, pages 948–956, 2017.

[48] 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 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4551–4560, 2019.