Rolling Shutter Correction with Intermediate Distortion Flow Estimation

Mingdeng Cao[1] Sidi Yang[2] Yujiu Yang[2] Yinqiang Zheng[1]✉
[1]The University of Tokyo [2]Tsinghua University

Abstract

This paper proposes to correct the rolling shutter (RS) distorted images by estimating the distortion flow from the global shutter (GS) to RS directly. Existing methods usually perform correction using the undistortion flow from the RS to GS. They initially predict the flow from consecutive RS frames, subsequently rescaling it as the displacement fields from the RS frame to the underlying GS image using time-dependent scaling factors. Following this, RS-aware forward warping is employed to convert the RS image into its GS counterpart. Nevertheless, this strategy is prone to two shortcomings. First, the undistortion flow estimation is rendered inaccurate by merely linear scaling the flow, due to the complex non-linear motion nature. Second, RS-aware forward warping often results in unavoidable artifacts. To address these limitations, we introduce a new framework that directly estimates the distortion flow and rectifies the RS image with the backward warping operation. More specifically, we first propose a global correlation-based flow attention mechanism to estimate the initial distortion flow and GS feature jointly, which are then refined by the following coarse-to-fine decoder layers. Additionally, a multi-distortion flow prediction strategy is integrated to mitigate the issue of inaccurate flow estimation further. Experimental results validate the effectiveness of the proposed method, which outperforms state-of-the-art approaches on various benchmarks while maintaining high efficiency. The project is available at https://github.com/ljzycmd/DFRSC.

1. Introduction

We often encounter distorted images/videos when relative movements occur between the scene and the camera during the acquisition process. For instance, a straight building may appear slanted in the captured photograph, while the blades of a flying helicopter may seem distorted. This phenomenon is generally referred to as the wobble or the "Jello effect.", which is caused by the rolling shutter (RS) mechanism of cameras. The image pixels are exposed from the top to the bottom sequentially, instead of capturing the entire frame all at once as in a global shutter (GS) camera. This RS mechanism is employed in CMOS sensors, which govern the cameras (*e.g.*, smartphones, digital cameras) in the consumer market, owing to their fast imaging and low cost. However, some unintended distortions would occur in the image content when capturing moving scenes, affecting our visual perception and deteriorating the performance of downstream tasks, especially 3D vision tasks [3, 5, 16, 23]. Consequently, developing effective and robust image/video RS correction (RSC) algorithms to remove such distortions holds significant research and practical application value.

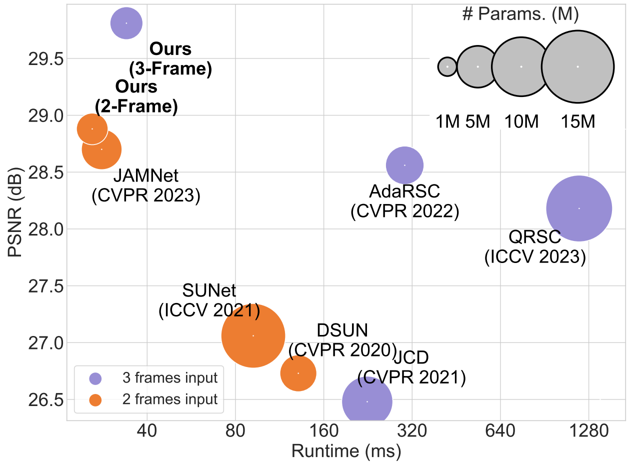


Figure 1. Model comparison in terms of PSNR (dB), runtime, and model size. The PSNR and runtime are calculated on the Fastec-RS [24] dataset with a resolution of 640 × 480 using an RTX 3090 GPU. The proposed method outperforms the state-of-the-art rolling shutter correction methods with higher efficiency.

To recover the latent distortion-free GS image corresponding to a specific exposure scanline of the RS image, previous research efforts [1, 8, 26, 29] have attempted to directly restore the underlying GS image from a single RS image by employing additional geometric constraints and priors. However, this kind of single-image-based approach is highly challenging and has limited effectiveness, as the motion states that form the distortion are unknown and strongly ambiguous, making the removal of such distortions from a single image highly ill-posed. Utilizing multiple consecutive images can significantly alleviate this issue by extracting the inter-frame motion information, thereby achieving better and more robust results [11, 21, 34, 41, 42], especially with deep-learning-based techniques [2, 7, 10, 24, 27]. Generally, to obtain the displacement fields (*i.e*., the correction fields) from RS to GS images, these methods usually first estimate the inter-frame motion field (optical flow) between the RS frames and then use the relationship [6, 9, 27, 42] between RS and GS images to transform it into a correction field, thus obtaining the desired GS images through image warping operations. These methods have achieved great success but struggle when faced with non-linear and large motion due to the following reasons: **1) First,** some methods [6, 9, 25, 27] employ off-the-shelf motion modeling networks to estimate the inter-frame motion of RS images. However, since these optical flow estimation networks have not been trained on RS videos, their estimation results may contain distortions and exhibit erroneous dynamic behavior, making it difficult to obtain an accurate correction field for recovering underlying GS images. **2) Alternatively**, other methods [2, 7, 24, 40] estimate the inter-frame optical flow within the RSC model and are trained with RS images/videos, typically using local correlation [4], which makes it difficult to model large motions. **3) Moreover**, to obtain the correction field, the estimated inter-frame optical flow needs to be further linearly scaled based on the constant velocity assumption [2, 10, 24]. However, the motion in the real world is highly non-linear, rendering the obtained correction field inaccurate. Although the recent work [27] proposes estimating a quadratic correction field, the motion in real-world scenarios is often more complex than quadratic and is thus more challenging to model.

To move beyond these limitations, in this paper, we propose to directly estimate the correction field from GS to RS images, dubbed *Distortion Flow* [1]. More specifically, we first generate the latent GS feature based on the extracted RS features. Then, we obtain an initial estimation of the distortion flow by establishing the global correlation between the RS and GS features, with the proposed flow attention mechanism. The GS feature and flow are continuously refined through a coarse-to-fine decoder, which fuses the warped RS appearance information to the GS feature and updates the flow. Simultaneously, we integrate a multi-distortion field decoding strategy to further alleviate the occlusion problem. The RS features are backwardly warped into the GS counterparts using multiple distortion fields and decoded along with the GS features to generate the final GS image. As shown in Fig. 1, our method achieves highly competitive results on various datasets more efficiently.

Our contributions are threefold and can be summarized as follows. **1)** We propose a novel framework for the RSC task that directly predicts *Distortion Flow* from consecutive RS frames to recover the underlying GS frame. **2)** We design a global correlation-based flow attention mechanism for GS feature and flow prediction, facilitating large motion prediction. In conjunction, a multi-distortion flow prediction strategy is formulated to further improve the performance. **3)** Extensive experiments demonstrate that the proposed method achieves substantial performance improvements against state-of-the-art methods on multiple datasets while maintaining higher efficiency.

2. Related Work

2.1. Deep Rolling Shutter Correction

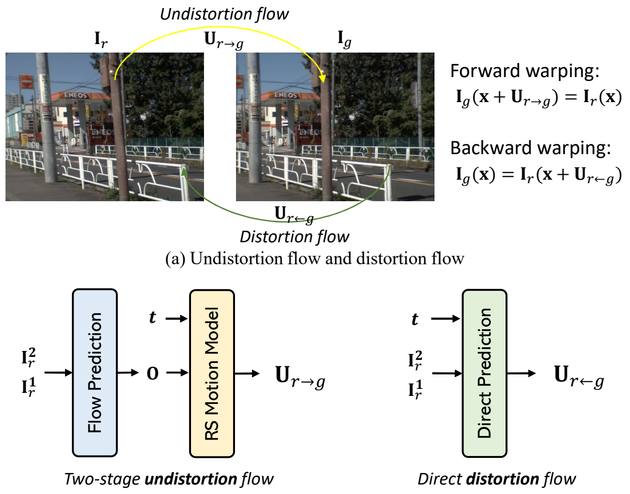
Existing works of RSC fall into two categories: single-image-based and multi-frame-based methods. For the former, previous methods apply different geometric assumptions, such as straight lines kept straight [28], vanishing direction restraint [26], and analytical 3D straight line RS projection model [21]. Driven by the surge of deep learning, the first learning-based model proposed in [30] attempts to remove RS distortions from a single distorted image. However, single-image-based models often exhibit unsatisfactory performance due to their reliance on either strong assumptions or inconspicuous features. This limitation hinders their ability to accurately capture the complexity of the underlying data, leading to sub-optimal results.

To tackle these limitations, multi-frame-based methods are adopted to model the RS motion, which can be categorized into classical and learning-based models. For the classical methods, modeling the RS motion from the uncalibrated RS images and two consecutive frames are respectively studied in [11, 20] and [34, 41]. For the learning-based methods, the works [7, 24] are proposed to model the RS motions between two consecutive RS frames by constructing cost volumes. Considering the blur in RS images, Zhong *et al*. [40] further designed a three-frame-based model to remove the blur and RS distortion simultaneously. To alleviate the inaccurate displacement field estimation and warping, Cao *et al*. [2] proposed to predict multiple fields and warp the RS features adaptively. Fan *et al*. [10] and Qu *et al*. [27] proposed a joint motion and appearance modeling network and a quadratic RS motion solver, respectively, achieving new heights.

2.2. Inter-frame Motion Modeling

To model the motions across frames, computing the matching cost volume to obtain the correspondence is a classical way. Optical flow networks, such as [4, 13, 32, 33], usually apply local correlations to obtain the final flow in a coarse-to-fine strategy with efficiency. RAFT [33] proposes an all-pair correlation volume and designs a recur-

[1] Distinguished with the *undistortion flow* field from RS to GS, used in [6, 7, 9, 24, 27, 40].



(a) Undistortion flow and distortion flow

(b) Two-stage and direct correction flow prediction

Figure 2. (a) Schematic of the undistortion flow and distortion flow. (b) Comparison between the two-stage flow estimation and the proposed direct distortion flow estimation.

rent strategy to refine the predicted flow continuously, obtaining considerable accuracy improvement. GMFlow [38] also constructs a global correlation between two frames to aggregate the coordinate grid as the correspondence, forming a new paradigm for optical flow estimation. These models are widely employed as off-the-shelf modules for motion modeling in many video-related tasks, like video frame interpolation, video enhancement, video editing, and RSC task [6, 9, 25, 27]. However, since these methods are not trained in the specific domain data, the motion estimation in the other tasks is inaccurate. To this point, Super-Slomo [14] introduces a mask to handle the occlusion explicitly and provides a standard formulation for synthesizing intermediate frames. RIFE [12] and IFRNet [18] propose task-oriented flow distillation losses to provide a prior intermediate flow in training. AMT [22] further adapts the all-pair correlation for efficient frame interpolation. In this paper, we also propose to learn the distortion flow for the RSC task, by constructing global correlations between the underlying GS image and input RS frames.

3. Proposed Method

3.1. Preliminary

RS cameras expose the pixels in a row-by-row manner, and each scanline has a different timestamp and motion. Thus the RS image $\mathbf{I}_r$ can be formed by row-by-row stacking the virtual GS images corresponding to each row timestamp:

$$[\mathbf{I}_r(\mathbf{x})]_i = [\mathbf{I}_g^i(\mathbf{x})]_i, \quad 0 \le i \le H-1, \tag{1}$$

where $\mathbf{I}_g^i$ is the virtual GS image corresponding to the timestamp of i-th RS image row, $[\cdot]_i$ is the operation to extract the i-th image row, H and $\mathbf{x}$ are respectively the image height and the pixel location. More generally, we can obtain the j-th RS image row with the displacement field $\mathbf{u}_{i\to j}$ from the j-th row of the RS image to the i-th virtual GS image:

$$[\mathbf{I}_r(\mathbf{x})]_j = [\mathbf{I}_g^i(\mathbf{x}+\mathbf{u}_{j\to i})]_j, \quad 0 \le i, j \le H-1. \tag{2}$$

With the above equation, we can obtain the *RS undistortion flow* field (the yellow line in Fig. 2(a)) $\mathbf{U}_{r\to i}$ from the RS image to the i-th virtual GS image, by stacking all $\mathbf{u}_{j\to i}$ from $j = 0$ to $j = H-1$. Therefore, we can recover the i-th underlying GS image (usually corresponding to the first scanline [7], and the middle scanline [2, 10, 24, 27] of the RS image) by estimating the undistortion flow field and using a forward warping operation like the differential forward warping (DFW) module [24]. As shown in the left part of Fig. 2(b), the velocity of the RS image pixels (approximated as the optical flow between consecutive RS frames) is first estimated, then the undistortion flow can be usually calculated by rescaling the flow under the constant velocity assumption [2, 6, 9, 10, 24]. However, accurate $\mathbf{U}_{r\to i}$ is hard to estimate with such a linear model since the motion in the real world is highly complex and non-linear, even with a recently proposed quadratic motion solver [27]. Moreover, the inaccurate $\mathbf{U}_{r\to i}$ further results in undesired warping artifacts with the DFW module, *e.g.*, black holes shown in [6].

In contrast to the Eq. 2 that forms the RS image from a sequence of GS images, it is feasible to derive the underlying i-th GS image from an RS image, when with the motion displacement field $\mathbf{U}_{r\leftarrow i}$ from GS to RS images (the green line in Fig. 2(a)):

$$\mathbf{I}_g^i(\mathbf{x}) = \mathbf{I}_r(\mathbf{x}+\mathbf{U}_{r\leftarrow i}(\mathbf{x})). \tag{3}$$

Thus we can obtain the underlying GS image by sampling pixels in the RS image with interpolation operations, *e.g.*, bilinear, and bicubic. Since the motion field attributes distort the GS image into the RS image, we dub it *distortion flow* field. However, the challenge arises since the intermediate GS image is unknown.

In this work, we propose to estimate the intermediate distortion flow from the underlying desired GS image to the RS image in a single-stage manner, as depicted in the right part of Fig. 2(b). Note that the time $\mathbf{t}$ determines the recovery of a specific GS image, and is optional for the RSC task since we aim to recover only one GS frame corresponding to a specific scanline (first or middle) of the RS frame.

3.2. Model Overview

Our method aims to alleviate the inaccurate motion modeling under large and complex non-linear motions in the RSC task, by directly estimating the intermediate distortion flows. Our method takes N consecutive RS frames

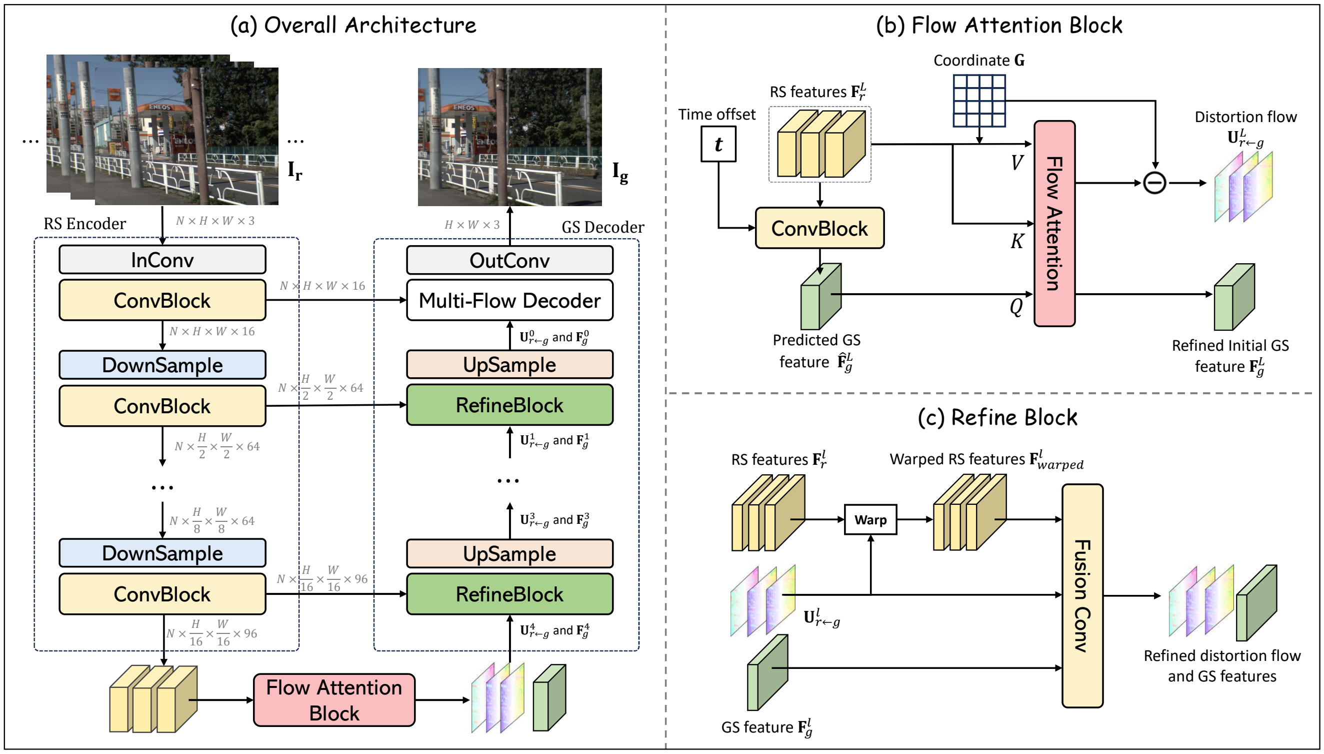

(a) Overall Architecture
...
...
$\mathbf{I_r}$
$\mathbf{I_g}$
RS Encoder
$N \times H \times W \times 3$
$H \times W \times 3$
GS Decoder
InConv
ConvBlock
$N \times H \times W \times 16$
$N \times H \times W \times 16$
DownSample
ConvBlock
$N \times \frac{H}{2} \times \frac{W}{2} \times 64$
$N \times \frac{H}{2} \times \frac{W}{2} \times 64$
...
$N \times \frac{H}{8} \times \frac{W}{8} \times 64$
DownSample
ConvBlock
$N \times \frac{H}{16} \times \frac{W}{16} \times 96$
$N \times \frac{H}{16} \times \frac{W}{16} \times 96$
OutConv
Multi-Flow Decoder
$\mathbf{U}^0_{r\leftarrow g}$ and $\mathbf{F}^0_g$
UpSample
RefineBlock
$\mathbf{U}^1_{r\leftarrow g}$ and $\mathbf{F}^1_g$
...
$\mathbf{U}^3_{r\leftarrow g}$ and $\mathbf{F}^3_g$
UpSample
RefineBlock
$\mathbf{U}^4_{r\leftarrow g}$ and $\mathbf{F}^4_g$
Flow Attention Block
(b) Flow Attention Block
Coordinate $\mathbf{G}$
RS features $\mathbf{F}^L_r$
Time offset
$\boldsymbol{t}$
ConvBlock
V
K
Q
Flow Attention
Distortion flow $\mathbf{U}^L_{r\leftarrow g}$
Predicted GS feature $\hat{\mathbf{F}}^L_g$
Refined Initial GS feature $\mathbf{F}^L_g$
(c) Refine Block
RS features $\mathbf{F}^l_r$
Warped RS features $\mathbf{F}^l_{warped}$
Warp
$\mathbf{U}^l_{r\leftarrow g}$
Fusion Conv
Refined distortion flow and GS features
GS feature $\mathbf{F}^l_g$

Figure 3. Overview of the proposed method (a) and the detailed architecture of the key components (b), (c). Our model directly predicts the distortion flow for efficient and high-quality RSC.

as input, and recovers the latent GS image corresponding to the timestamp of the middle scanline of the middle input RS frame, consistent with the settings in previous works [2, 10, 24, 27, 40]). The overall architecture of the proposed method is illustrated in Fig. 3. We first extract multi-scale frame-level RS features, using a weight-sharing image encoder. After that, we obtain the initial distortion flow along with the GS features at the lowest resolution with a global correlation-based flow attention mechanism. Then, the coarse-to-fine decoder refines and upscales the resolution of the flow and GS features simultaneously. The final GS image is obtained by a multi-flow predicting strategy.

3.3. Intermediate Distortion Flow Estimation

Initial distortion flow estimation. After obtaining the L-scale features $\{\mathbf{F}^l\}_{l=0}^{L}$ of the input N RS frames $\mathbf{I}_r \in \mathbb{R}^{N\times H\times W\times 3}$ from the encoder, we can directly predict the initial intermediate distortion flow $\mathbf{U}_{r\leftarrow g}$ with the lowest resolution features $\mathbf{F}^L \in \mathbb{R}^{N\times H'\times W'\times D}$ in a naive way:

$$\mathbf{U}^L_{r\leftarrow g} = \text{IDFE}(\mathbf{F}^L, \mathbf{t}), \tag{4}$$

where IDFE is the prediction network, and $\mathbf{t}$ is the exposure time offset between the target GS image and the middle scanline of the RS frame.

To obtain a more accurate intermediate distortion flow estimation under large motions, we further perform global correlation modeling across the desired underlying GS and RS features. The Attention $=$ Softmax$(\frac{\mathbf{QK}^T}{\sqrt{d}})\mathbf{V}$ mechanism [35] tries to aggregate the value $V \in \mathbb{R}^{S\times d}$ with the correlation between the query $Q \in \mathbb{R}^{S\times d}$ and value $K \in \mathbb{R}^{S\times d}$, excelling at long-range modeling and correlation modeling. We extend such an operation to build the global correlation between GS and RS frames for distortion flow estimation and RS feature warping shown in Fig. 3(b), dubbed flow attention. While the GS feature is missing, we thus estimate it $\mathbf{F}^L_g \in \mathbb{R}^{H'\times W'\times D}$ firstly by fusing the consecutive RS features with the time offsets condition:

$$\mathbf{F}^L_g = \text{ConvBlock}(\mathbf{F}^L, \mathbf{t}). \tag{5}$$

Let the GS features and RS features serve as *query* and *key*, respectively, and we can compute the attention map between them:

$$\mathbf{M} = \text{Softmax}\left(\frac{\mathbf{F}^L_g\mathbf{F}^L}{\sqrt{D}}\right) \in \mathbb{R}^{N\times H'\times W'\times H'\times W'}, \tag{6}$$

where each element (n, i, j, k, l) in $\mathbf{M}$ represents the correspondence probability between the GS feature $\mathbf{F}^L_g(i, j)$ and the RS frame feature $\mathbf{F}^L(n, k, l)$. Note that the above equation is consistent with the differentiable matching layer [36, 38] in image matching and optical flow estimation. With the global correlation matrix $\mathbf{M}$, we can simultaneously compute the globally warped RS features and the distortion flow by aggregating 1) the RS features and 2) the 2D coordinates grid $\mathbf{G} \in \mathbb{R}^{H' \times W' \times 2}$ of the RS frame, respectively. As a result, both $\mathbf{F}^L$ and $\mathbf{G}$ serve as the *value* for the decoding:

$$\mathbf{F}^L_{warped} = \mathbf{M}\mathbf{F}^L \in \mathbb{R}^{N \times H' \times W' \times D}, \tag{7}$$

$$\mathbf{U}^L_{r \leftarrow g} = \mathbf{M}\mathbf{G} - \mathbf{G} \in \mathbb{R}^{N \times H' \times W' \times 2}. \tag{8}$$

The warped RS features are further used to refine the predicted $\mathbf{F}^L_g$ using a convolutional block.

Employing the global correlation attention mechanism enables obtaining a more precise distortion flow from RS frames, especially with non-linear and large motions. In addition, the predicted GS feature can be further refined with the globally warped RS features, by fusing the complementary information RS features.

Progressive refinement. The predicted initial distortion flow and GS feature at the lowest resolution are progressively refined by the decoder. Inspired by [10], we employ the joint appearance and motion refinement strategy, while we directly predict the upsampled refined distortion flow rather than scale the optical flow between RS frames. Specifically, given the current refined distortion flow $\mathbf{U}^l_{r \leftarrow g}$ and the GS features $\mathbf{G}^l_g$ at level l, we first warp the RS features extracted from the image encoder at the corresponding level to the GS candidates:

$$\mathbf{F}^l_{warped} = \mathcal{W}(\mathbf{F}^l, \mathbf{U}^l_{r \leftarrow g}), \tag{9}$$

where $\mathcal{W}$ is the backward warping operation. Next, the warped RS features, distortion flow, and GS feature are fused and upscaled to the refined distortion flow and GS feature at the next scale $l-1$:

$$\mathbf{U}^{l-1}_{r \leftarrow g}, \mathbf{F}^{l-1}_g = \text{Upsample}(\text{FusionBlock}(\mathbf{U}^l_{r \leftarrow g}, \mathbf{F}^l_g, \mathbf{F}^l_{warped})). \tag{10}$$

By progressively fusing the complementary information from the RS features, more accurate distortion flow and corresponding GS features are obtained to be decoded as the final GS image.

Multi-distortion flow fields decoding. At the 0-level with the largest resolution, we further employ a multiple distortion fields prediction strategy [2] to alleviate some incorrectly estimated displacement in the distortion flow. With the refined $\mathbf{U}^0_{r \leftarrow g}$ and $\mathbf{F}^0_g$, rather than utilize them to synthesize the GS image directly, we instead predict multiple groups of fields with a convolutional block:

$$\{\mathbf{U}_{1,r \leftarrow g}, \cdots, \mathbf{U}_{G,r \leftarrow g}\} = \text{ConvBlock}(\mathbf{U}_{r \leftarrow g}, \mathbf{F}_g, \mathbf{F}_{warped})). \tag{11}$$

Therefore, the RS features $\mathbf{F}^0$ are further warped according to Eq. 9, resulting in G groups of warped RS features. The final GS image $\mathbf{I}_g$ is predicted from the refined GS features, undistortion flow, and warped multiple RS features.

3.4. Training Strategy

GS image supervision. Following previous works [2, 7, 10, 24, 40], we employ a combination of the Charbonnier loss [19]

$$\mathcal{L}_c = d(\mathbf{I}_{gt} - \mathbf{I}_g) \tag{12}$$

and the Perceptual loss [15]

$$\mathcal{L}_p = \|\phi(\mathbf{I}_{gt}) - \phi(\mathbf{I}_g)\|_1, \tag{13}$$

for the recovered GS image supervision, where $d(x) = \sqrt{(x)^2 + \epsilon^2}$ is a distance function, and ϵ is set to $1e^{-3}$, and ϕ is the extractor to obtain the features from layer `Conv5_4` of pretrained VGG-19 network [31].

Distortion flow supervision. To ensure the accuracy of the estimated distortion flow, we employ an indirect supervision method that ensures the backward warped RS images with the undistortion flow align consistently with the ground truth GS image:

$$\mathcal{L}_w = \frac{1}{L} \sum_{l=0}^{L} d(\mathbf{I}^{l,warped}_r - \mathbf{I}^l_{gt}), \tag{14}$$

where $\mathbf{I}^{l,warped}_r = \mathcal{W}(\mathbf{I}^l_r, \mathbf{U}^l_{r \leftarrow g})$ is the warped downsampled RS frames $\mathbf{I}_r$ at the l-th level, and $\mathbf{I}^l_{gt}$ is the downsampled ground truth GS image at level l.

The total loss for the model training can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_c + \lambda_1 \mathcal{L}_p + \lambda_2 \mathcal{L}_w, \tag{15}$$

where λ_1 and λ_2 are loss weights.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate the proposed method on both synthetic datasets Fastec-RS [24], Carla-RS [24], and the real-world datasets BS-RSC [2]. The Fastec-RS dataset is synthesized from the extremely high-speed videos captured by a GS camera, mainly containing RS effects caused by horizontal camera movements. Another synthetic dataset Carla-RS is generated from a virtual 3D environment, with constant translational velocity and angular rate during the RS video sequence generation process. The recently proposed BS-RSC dataset is collected from the real world. The RS videos and corresponding GS videos are captured simultaneously by a well-designed beam-splitter acquisition system. The scenes contain natural non-linear and large motions, including both camera and objects.

Implementation details. During the training process, our model accepts $N = 3$ consecutive RS frames in RGB format as input, while we also train a 2-frame-based model

Method	# Params.	Runtime	# NF	Fastec-RS			Carla-RS		
	(Million)	(ms)		PSNR↑(dB)	SSIM↑	LPIPS↓	PSNR↑(dB)	SSIM↑	LPIPS↓
DiffSfM [42]	-	4.7×10^5	2	21.44	0.710	0.2180	21.28	0.775	0.1322
DSUN [24]	3.91	131	2	26.73	0.819	0.0995	26.46	0.807	0.0703
SUNet [7]	12.0	92	2	27.06	0.825	0.1030	29.18	0.850	0.0658
VideoRS [25]	24.26	1.3×10^6	2	28.57	0.844	-	**31.43**	0.919	-
JAMNet [10]	4.73	28	2	28.70	0.865	**0.0691**	30.70	0.905	0.0371
Ours (2F)	**2.87**	**26**	2	**28.88**	**0.870**	0.0699	31.33	**0.921**	**0.0228**
JCD [40]	7.51	225	3	26.48	0.821	0.0943	27.75	0.836	0.0595
AdaRSC [2]	4.25	302	3	28.56	0.855	0.0796	-	-	-
QRSC (3F) [27]	12.72	401	3	28.18	0.853	0.0912	29.81	0.919	0.0313
QRSC (4F) [27]	12.74	759	4	28.26	0.854	0.0901	30.98	0.925	0.0282
QRSC (5F) [27]	12.75	1149	5	29.49	0.872	0.0814	32.01	**0.933**	0.0253
Ours (3F)	**3.15**	**34**	3	**30.00**	**0.882**	**0.0665**	**32.10**	0.930	**0.0218**

Table 1. Quantitative comparison against the state-of-the-art methods on the synthetic RSC datasets Carla-RS [24] and Fastec-RS [24]. Our method achieves highly competitive results while maintaining high efficiency. #NF indicates the input RS frames of the model. The runtime is calculated using an NVIDIA RTX 3090 GPU.

Method	BS-RSC		ACC	
	PSNR↑(dB)	SSIM↑	PSNR↑(dB)	SSIM↑
DiffSfm [42]	19.80	0.698	15.74	0.551
DSUN [24]	25.21	0.833	22.39	0.780
SUNet [7]	27.76	0.875	27.29	0.870
JAMNet [10]	32.93	0.941	32.71	0.940
Ours (2F)	**33.39**	**0.947**	**33.21**	**0.947**
JCD [40]	25.59	0.841	23.73	0.808
AdaRSC [2]	28.23	0.882	28.73	0.892
QRSC (5F) [27]	33.50	0.946	33.36	0.945
Ours (3F)	**34.48**	**0.954**	**34.35**	**0.954**

Table 2. Quantitative comparison against the state-of-the-art methods on the real-world RSC dataset BS-RSC [2].

that inputs two frames. The feature scales $L = 4$. For the data augmentation, the input RS frames are first randomly cropped with a width of 256 while keeping the height unchanged, and a random horizontal flip is performed on the cropped patch. The loss hyper-parameters are set to $\lambda_1 = 0.005$, $\lambda_2 = 0.05$. The model is trained for 150k iterations with a step learning rate adjustment strategy. When testing, no augmentation is applied to the input consecutive RS frames. The experiments are conducted on the PyTorch platform on a single NVIDIA V100 GPU. The initial learning rate is set to 4×10^{-4}, and the ADAM optimizer [17] is employed to update the model parameters.

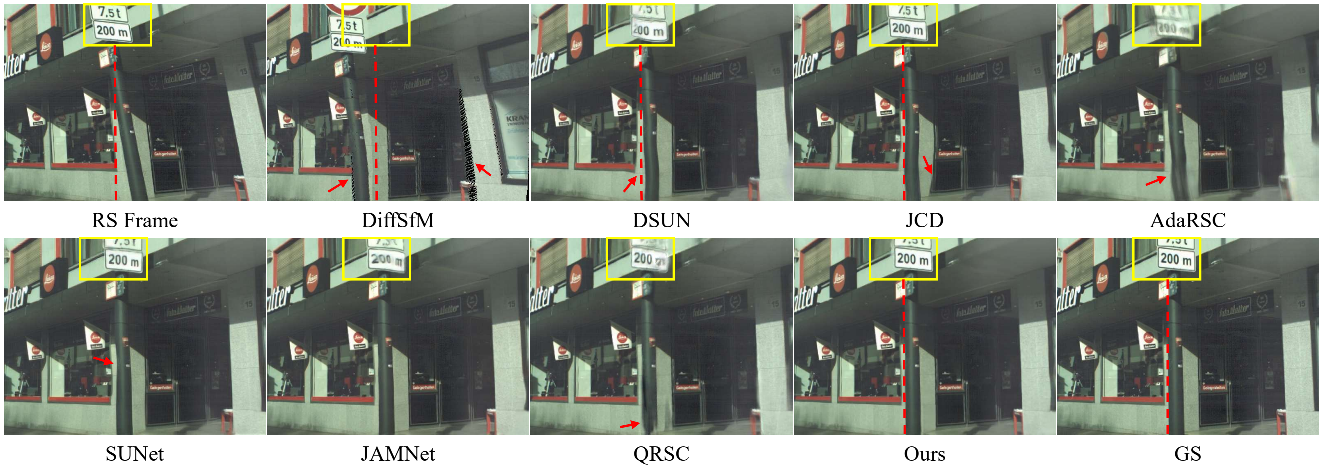
Evaluation metrics. Both PSNR and SSIM [37] are employed to evaluate the correction accuracy quantitatively. Meanwhile, the learned perceptual metric LPIPS [39] is also applied to measure the visual quality quantitatively. In addition, the corrected RS frames are also displayed for the qualitative comparison.

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

We compare the proposed method to the state-of-the-art RSC methods quantitatively and qualitatively, including **1)** traditional method DiffSfM [42], **2)** deep learning-based methods DSUN [24], SUNet [7], VideoRS [25], JAMnet [10] that take two consecutive frames as input, and **3)** deep learning-based methods JCD [40], AdaRSC [2], QRSC [27] that require inputting three or more frames. We also implemented two versions of the proposed method: 2-frame-based and 3-frame-based, to better demonstrate the effectiveness of the proposed method.

Quantitative comparison. Table 1 presents the performance of different methods on the synthetic datasets Factec-RS and Carla-RS. We see that the proposed method achieved highly competitive performance that obtains higher PSNR, SSIM, and lower LPIPS than the state-of-the-art methods JAMNet [10] and QRSC [27], thanks to the direct distortion flow estimation strategy and the model design. The quantitative results on the real-world BS-RSC dataset are shown in Tab. 2, where the proposed method shows significant improvements against other methods. Specifically, our 3-frame-based model achieves 2.5dB PSNR improvement compared to the 3-frame-based QRSC (3F), and even surpasses the 5-frame-based QRSC (5F) with about 1dB PSNR. Note that BS-RSC contains both camera and object motions in the real world.

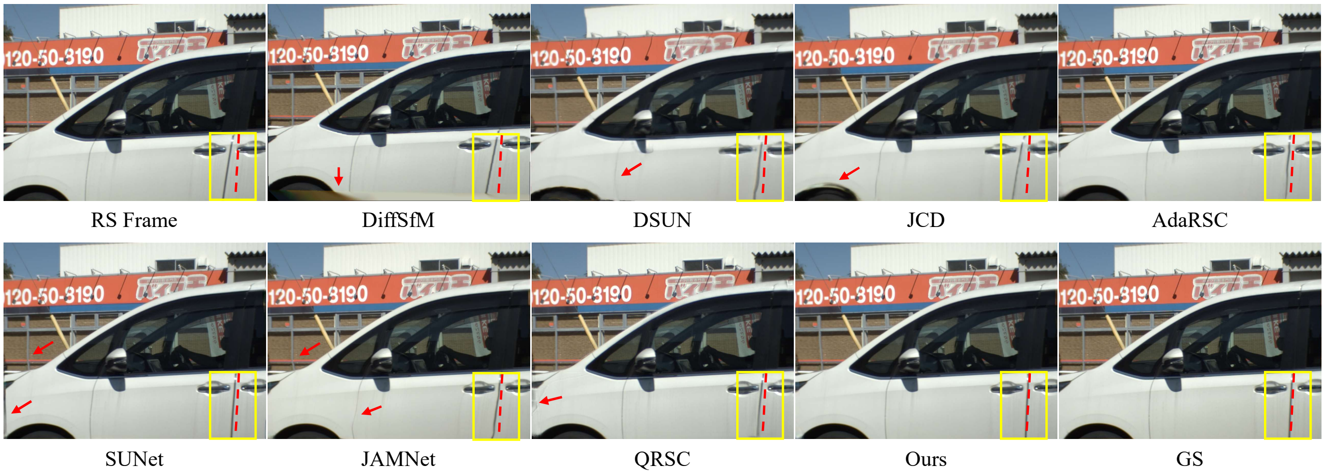
These competitive results demonstrate the effectiveness of the proposed method in removing the RS effects under non-linear and large motions. Unlike previous methods that estimate the optical flows between RS frames and utilize linear [2, 10, 24] or quadratic [27] motion models to obtain the correction fields, our methods directly predict the dis-



RS Frame DiffSfM DSUN JCD AdaRSC

SUNet JAMNet QRSC Ours GS

Figure 4. Qualitative results comparison against state-of-the-art methods on the synthetic Fastec-RS dataset [24]. Our method removes the RS distortions well and preserves more details in the recovered GS image on such an occluded scene.



RS Frame DiffSfM DSUN JCD AdaRSC

SUNet JAMNet QRSC Ours GS

Figure 5. Qualitative results comparison against state-of-the-art methods on the real-world BS-RSC dataset [2]. Our method is effective and robust in recovering the latent GS image accurately from the RS frames distorted by complex non-linear and large motions.

tortion flow and can better model the non-linear motions to obtain better results. Meanwhile, as shown in the right part of Tab. 2, our method also achieved superior results on the ACC dataset [27], which is derived from BS-RSC by excluding frames with constant motions. These quantitative results demonstrate that the proposed method is effective and robust in removing RS distortions under complex non-linear and large motions.

Qualitative comparison. Figures 4 and 5 illustrate the qualitative results of different methods on the synthetic dataset Fastec-RS and the real-world dataset BS-RSC, respectively. As for the occluded scene shown in Fig. 4, we see that existing methods struggle to either remove the distortions or preserve details for high-quality GS restoration. In contrast, our method successfully recovers the corrected GS image while preserving more details (*e.g.*, the number marked with the yellow box). Meanwhile, as for the scene containing both non-linear camera motion and object motion shown in Fig. 5, existing methods can hardly obtain the correct shape of the latent GS image (marked by the yellow box). These methods make it difficult to obtain an accurate correction field with a linear motion model, *e.g.*, DSUN, AdaRSC, and JAMNet, even with a quadratic motion model, *i.e.*, QRSC. Thanks to the proposed direct intermediate distortion flow estimation and the network design, our model performs better in removing the RS distortions caused by complex and large motions and recovering the desired GS image accurately.

Efficiency comparison. As shown in Tab. 1, our method is also highly competitive in terms of efficiency. More specif-

Model	PSNR	SSIM	# Params.
W/o motion modeling	26.88	0.833	2.43
W/ undistortion flow $\mathbf{U}_{r\rightarrow g}$	27.78	0.847	2.86
W/ distortion flow $\mathbf{U}_{r\leftarrow g}$	28.39	0.864	2.79
Full model	28.88	0.870	2.87

(a) Effectiveness of the distortion flow estimation.

Model	PSNR	SSIM	# Params.
W/o Flow attention	28.52	0.865	2.79
W/ 1 field	28.73	0.867	2.87
W/ 4 fields	28.88	0.870	2.87
W/ 8 fields	28.82	0.865	2.88

(b) Ablation study on the model design.

Table 3. Ablation study of the motion modeling and the model design. The settings employed in our final model are highlighted.

ically, our 2-frame-based model achieves a higher PSNR than the previous most efficient RSC model JAMNet, and has 40% fewer parameters. Moreover, our method (3-frame-based version) realizes a significant performance gain on all three datasets while only slightly slower than JAMNet. Compared to QRSC, the proposed method is more than 30 times faster with a much smaller number of model parameters. This is because QRSC requires computing the optical flows several times among input RS frames using the off-the-shelf flow models, while our method performs RSC in an end-to-end manner.

4.3. Ablation Studies

We ablate the proposed method in terms of the distortion flow estimation and the network modules with our 2-frame-based model on the popular Fastec-RS dataset, and the ablation of our 3-frame-based model on the real-world BS-RSC dataset can be found in the supplementary materials.

Distortion flow estimation. To validate the effectiveness of the direct distortion flow estimation for the RSC task, we first remove the flow-attention module and multi-distortion flow decoding module, then **1)** remove the flow estimation and warping operation to obtain a vanilla encoder-decoder-like model (without motion modeling), **2)** replace the distortion flow estimation with direct undistortion flow $\mathbf{U}_{r\rightarrow g}$ estimation and apply differential forward warping [24] module for warping. The results of the above model variants are shown in Tab. 3a. We see that a vanilla encoder-decoder model achieves the lowest metrics, while the models with motion modeling (with undistortion or distortion flow) significantly improve the performance. This verifies that inter-frame motion modeling is beneficial and necessary to achieve high-quality RSC results. Meanwhile, the model with distortion flow estimation obtains higher PSNR and SSIM than the undistortion flow-based model. We argue that the backward warping operation contributes to the performance improvement. In addition, when the distortion flow-based model with the delicate network design (*i.e.*, the global correlation-based flow attention and multi-distortion flow decoding strategy), the performance has been further improved with a slight parameter number increase.

Flow attention and multi-flow decoder. As shown in Tab. 3b, when adding the flow attention module or the multi-distortion flow decoding, PSNR and SSIM metrics have been further improved by better modeling the large complex motions and occlusions. Meanwhile, as the flow group number increases, the model exhibits minor performance fluctuations. However, it still achieves improvement when compared to the single-field-based model.

The Number of input RS frames. As shown in Tabs. 1 and 2, the performance of the two-frame-based version model declines drastically on all datasets. With two frames, some contents in the latent GS image corresponding to the middle scanline of the second RS image still cannot be found in the first RS image. As a result, the missing regions should be generated, which is highly challenging. When with three frames, more complementary appearance information in the neighboring RS frames can be aggregated to obtain a higher-quality GS image.

5. Limitation

Although our method achieves highly competitive performance in recovering high-quality GS images and surpasses existing methods with a large margin, it still can not fully address the RS distortions encountered in real-world scenarios, constrained by the scale of existing datasets and the unknown camera parameters of the capture. In the following work, we want to integrate explicit camera exposure parameters into the model design for a more effective and generalized real-world RSC task.

6. Conclusion

This paper explores the intermediate distortion flow estimation for the high-quality performance on the RSC task. A novel framework, equipped with a global correlation-based flow attention module and a multi-distortion flow decoding strategy, is proposed to estimate the distortion flows from the latent GS image to the RS frames directly. Experimental results on both synthetic and real-world datasets demonstrate the effectiveness of the proposed method, and that it can remove the RS distortions under complex non-linear and large motions efficiently. We hope the proposed method can serve as a new paradigm to develop more effective and efficient methods for the RSC task.

Acknowledgements. This research was supported in part by JSPS KAKENHI Grant Numbers 22H00529, 20H05951, JST-Mirai Program JPMJMI23G1.

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. 1

[2] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 17785–17793, 2022. 2, 3, 4, 5, 6, 7

[3] Yuchao Dai, Hongdong Li, and Laurent Kneip. Rolling shutter camera relative pose: Generalized epipolar geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4132–4140, 2016. 1

[4] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 2

[5] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 1

[6] Bin Fan and Yuchao Dai. Inverting a rolling shutter camera: bring rolling shutter images to high framerate global shutter video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4228–4237, 2021. 2, 3

[7] Bin Fan, Yuchao Dai, and Mingyi He. Sunet: symmetric undistortion network for rolling shutter correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4541–4550, 2021. 2, 3, 5, 6

[8] Bin Fan, Yuchao Dai, and Ke Wang. Rolling-shutter-stereo-aware motion estimation and image correction. *Computer Vision and Image Understanding*, 213:103296, 2021. 1

[9] Bin Fan, Yuchao Dai, Zhiyuan Zhang, Qi Liu, and Mingyi He. Context-aware video reconstruction for rolling shutter cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17572–17582, 2022. 2, 3

[10] Bin Fan, Yuxin Mao, Yuchao Dai, Zhexiong Wan, and Qi Liu. Joint appearance and motion learning for efficient rolling shutter correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5671–5681, 2023. 2, 3, 4, 5, 6

[11] Matthias Grundmann, Vivek Kwatra, Daniel Castro, and Irfan Essa. Calibration-free rolling shutter removal. In *IEEE International Conference on Computational Photography*, pages 1–8, 2012. 2

[12] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision*, pages 624–642. Springer, 2022. 3

[13] 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. 2

[14] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9000–9008, 2018. 3

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 5

[16] Jae-Hak Kim, Cesar Cadena, and Ian Reid. Direct semi-dense slam for rolling shutter cameras. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1308–1315. IEEE, 2016. 1

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[18] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang. Ifrnet: Intermediate feature refine network for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1969–1978, 2022. 3

[19] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 5

[20] 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*, 2018. 2

[21] 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. 2

[22] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023. 3

[23] Bangyan Liao, Delin Qu, Yifei Xue, Huiqing Zhang, and Yizhen Lao. Revisiting rolling shutter bundle adjustment: Toward accurate and fast solution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4863–4871, 2023. 1

[24] 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. 1, 2, 3, 4, 5, 6, 7, 8

[25] Eyal Naor, Itai Antebi, Shai Bagon, and Michal Irani. Combining internal and external constraints for unrolling shutter in videos. In *European Conference on Computer Vision*, pages 119–134. Springer, 2022. 2, 3, 6

[26] Pulak Purkait, Christopher Zach, and Ales Leonardis. Rolling shutter correction in manhattan world. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 882–890, 2017. 1, 2

[27] Delin Qu, Yizhen Lao, Zhigang Wang, Dong Wang, Bin Zhao, and Xuelong Li. Towards nonlinear-motion-aware and occlusion-robust rolling shutter correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10680–10688, 2023. 2, 3, 4, 6, 7

[28] 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*, 2016. 2

[29] 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. 1

[30] Vijay Rengarajan, Yogesh Balaji, and A. N. Rajagopalan. Unrolling the shutter: Cnn to correct motion distortions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[32] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2

[33] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2

[34] Subeesh Vasu, AN Rajagopalan, et al. Occlusion-aware rolling shutter rectification of 3d scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 636–645, 2018. 2

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[36] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 757–774. Springer, 2020. 4

[37] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 6

[38] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2022. 3, 5

[39] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[40] Zhihang Zhong, Yinqiang Zheng, and Imari Sato. Towards rolling shutter correction and deblurring in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9219–9228, 2021. 2, 4, 5, 6

[41] Bingbing Zhuang and Quoc-Huy Tran. Image stitching and rectification for hand-held cameras. In *European Conference on Computer Vision*, pages 243–260. Springer, 2020. 2

[42] 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. 2, 6