Context-Aware Video Reconstruction for Rolling Shutter Cameras

Bin Fan Yuchao Dai* Zhiyuan Zhang Qi Liu Mingyi He
School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China

Abstract

With the ubiquity of rolling shutter (RS) cameras, it is becoming increasingly attractive to recover the latent global shutter (GS) video from two consecutive RS frames, which also places a higher demand on realism. Existing solutions, using deep neural networks or optimization, achieve promising performance. However, these methods generate intermediate GS frames through image warping based on the RS model, which inevitably result in black holes and noticeable motion artifacts. In this paper, we alleviate these issues by proposing a context-aware GS video reconstruction architecture. It facilitates the advantages such as occlusion reasoning, motion compensation, and temporal abstraction. Specifically, we first estimate the bilateral motion field so that the pixels of the two RS frames are warped to a common GS frame accordingly. Then, a refinement scheme is proposed to guide the GS frame synthesis along with bilateral occlusion masks to produce high-fidelity GS video frames at arbitrary times. Furthermore, we derive an approximated bilateral motion field model, which can serve as an alternative to provide a simple but effective GS frame initialization for related tasks. Experiments on synthetic and real data show that our approach achieves superior performance over state-of-the-art methods in terms of objective metrics and subjective visual quality. Code is available at https://github.com/GitCVfb/CVR.

1. Introduction

Many modern CMOS cameras equipped with rolling shutter (RS) dominate the consumer photography market due to their low cost and simplicity in design, and are also prevalent in the automotive sector and motion picture industry [16, 48, 52, 62]. Within this acquisition mode, pixels on the rolling shutter CMOS sensor plane are exposed from top to bottom in a row-by-row fashion with a constant inter-row delay. This leads to undesirable visual distortions called the RS effect (*e.g.* wobble, skew) in the presence of fast motion, which is a hindrance to scene understanding and a nuisance in photography. With the increased demand for high quality and high framerate video of consumer-grade devices (*e.g.* tablets, smartphones), video frame interpolation (VFI) has attracted increasing attention in the computer vision community. Unfortunately, despite the remarkable success, the currently existing VFI methods [2, 18, 38, 39, 56] implicitly assume that the camera employs a global shutter (GS) mechanism, *i.e.* all pixels are exposed simultaneously. They are therefore unable to produce satisfying in-between frames with rolling shutter video acquired by *e.g.* these devices in dynamic scenes or fast camera movements, resulting in RS artifacts remaining [9].

To address this problem, many RS correction methods [13, 17, 24, 43, 55, 63] have been actively studied to eliminate the RS effect. In analogy to VFI generating non-existent intermediate GS frames from two consecutive GS frames, recovering the latent intermediate GS frames from two consecutive RS frames, *e.g.* [10, 24, 61, 62], serves as a tractable

Figure 1. **GS video reconstruction example.** The left column shows two input consecutive RS images, and three ground-truth GS images at time 0, 0.5, and 1, respectively. Rows to the right show five GS frames (at times 0, 0.25, 0.5, 0.75, 1) extracted by [9] (top) and our method (below), followed by two corresponding zoom-in regions. The orange box represents occluded black holes and the red box indicates motion artifacts specific to moving objects. Our method recovers higher fidelity GS images due to contextual aggregation and motion enhancement. Note that the black image edges by our method are because they are not available in both RS frames (*cf*. blue circle). Best viewed on Screen.

*Y. Dai is the corresponding author (daiyuchao@gmail.com).

goal that overcomes the limited acquisition framerate and RS artifacts of commercial RS cameras. This is significantly challenging because the output GS frames must follow coherence both temporally and spatially. To this end, traditional methods [61, 62] are often based on the assumption of constant velocity or constant acceleration camera motion, which struggle to accurately reflect the real camera motion and scene geometry, resulting in the persistence of ghosting and unsmooth artifacts [9, 24]. Recent deep learning-based solutions have achieved impressive performance, but they typically can only recover one GS image corresponding to a particular scanline, such as the first [10] or central [24, 60] scanline, limiting their potentials for view transitions from RS to multiple-GS.

In this paper, we tackle the task of reviving and reliving all latent views of a scene as beheld by a virtual GS camera in the imaging interval of two consecutive RS frames. Therefore, we must jointly deal with VFI and RS correction tasks, *i.e*. interpolating smooth and trustworthy distortion-free video sequences. It is worth mentioning that the most relevant work to our task is [9], which is dedicated to the geometry-aware RS inversion by warping each RS frame to its corresponding virtual GS counterpart. Nevertheless, as illustrated in Fig. 1, the GS images recovered by [9] still suffers from two limitations:

- **Masses of black holes** (*cf*. orange box). This is a common issue for warping-based methods (*e.g*. [9, 44, 61–63]) due to the occlusion between the RS and GS images, leading to the possibility of permanent loss of some valuable image contents. To maintain visual consistency, a cropping operation is used to discard the resulting holes, but may degrade the visual experience.
- **Noticeable object-specific motion artifacts** (*cf*. red box). When recording dynamic scenes, the moving object violates the constant velocity motion assumption of RS cameras used in [9], resulting in its inability to accurately capture motion boundaries specific to moving objects. Thus severe motion artifacts are generated.

In contrast, we investigate contextual aggregation and motion enhancement based on the bilateral motion field (BMF) to alleviate these issues, which aims to synthesize crisp and pleasing GS video frames by occlusion reasoning and temporal abstraction. Specifically, we propose *CVR* (**C**ontext-aware **V**ideo **R**econstruction architecture), which consists of two stages to recover a faithful and coherent GS video sequence from two input consecutive RS images. In the first initialization stage, we adopt a motion interpretation module to estimate the initial bilateral motion field, which warps the two RS frames to a common GS version. We design two schemes to achieve this goal. One is based on [9] which requires a pre-trained encoder-decoder network; the other is our proposed approximation of [9], without resorting to a deep network. Also, we show that this simple approximation is able to provide a feasible solution for the initial prediction. Afterward, a second refinement stage is introduced to handle black holes and ambiguous misalignments caused by occlusions and object-specific motion patterns. As a result of exploiting bilateral motion residuals and occlusion masks, it can guide the subsequent GS frame synthesis to reason about complex motion profiles and occlusions. Furthermore, inspired by [10], we propose a contextual consistency constraint to effectively aggregate the contextual information, such that the unsmooth areas can be enhanced in an adaptive manner. Extensive experimental results demonstrate that our method surpasses the state-of-the-art (SOTA) methods by a large margin in removing RS artifacts. Meanwhile, our method is capable of generating high-fidelity GS videos.

The main contributions of this paper are three-fold:

1) We propose a simple yet effective bilateral motion field approximation model, which serves as a reliable initialization for GS frame refinement.
2) We develop a stable and efficient context-aware GS video reconstruction framework, which can reason about complex occlusions, motion patterns specific to objects, and temporal abstractions.
3) Experiments show that our method achieves SOTA results while maintaining an efficient network design.

2. Related Work

Video frame interpolation has been widely studied in recent years, which can be categorized into phase-based [31, 32], kernel-based [5, 28, 36], and flow-based [2, 18, 38, 49] methods. With the latest advances in optical flow estimation [7, 50, 51], the flow-based VFI methods have been actively studied to explicitly exploit motion information. After the seminal work [18], subsequent improvements are dedicated to better intermediate flow estimation on one hand, such as quadratic [56], rectified quadratic [26], and cubic [4] flow interpolations. Moreover, Bao *et al*. [2] strengthened the initial flow field using the predicted depth map via a depth-aware flow projection layer. Park *et al*. estimated a symmetric bilateral motion [38] to produce the intermediate flows directly, and they have recently developed an asymmetric bilateral motion model [39] to refine the intermediate frame. On the other hand, better refinement and fusion of details were focused on, including contextual warping [2, 33, 34], occlusion inference [3, 57], cycle constraints [27, 42] for more accurate frame synthesis, and softmax splatting [35] for more efficient forward warping, *etc*.

All of these VFI approaches work with a common assumption that the camera employs a GS mechanism. Hence, they are incapable of correctly synthesizing the in-between frames in the case of RS images. In this paper, we integrate

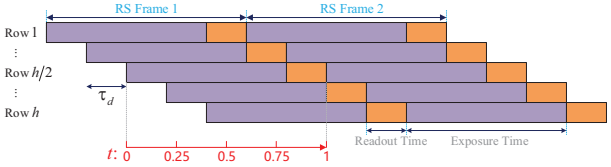

Figure 2. RS mechanism over two consecutive frames. We aim at recovering the latent GS images at time $t \in [0, 1]$.

an effective motion interpretation module to boost the reliable estimation of the initial flow field, yielding high-quality results without aliasing.

Rolling shutter correction advocates the mitigation or elimination of RS distortion, *i.e*. recovering the latent GS image, from a single frame [22, 43, 44, 63] or multiple frames [1, 15, 24, 47, 53, 61]. Dai *et al*. [6] derived the discrete two-view RS epipolar geometry. Zhuang *et al*. [61] proposed a differential RS epipolar constraint to undistort two consecutive RS images, whose stereo version was further explored in [12]. Likewise, Lao *et al*. [23] developed a discrete RS homography model to perform the plane-based RS correction. Zhuang and Tran [62] presented a differential RS homography to account for the scanline-varying poses of RS cameras. In addition, some additional assumptions are often taken into account, such as pure rotational motion [14, 22, 44, 45], Ackermann motion [40], and Manhattan world [41]. With the rise of deep learning, many appealing RS correction results have been achieved. For two input consecutive RS frames, Liu *et al*. [24] put forward a deep shutter unrolling network to estimate the latent GS frame, and Fan *et al*. [10] proposed a symmetric network architecture to efficiently aggregate the contextual cues. Zhong *et al*. [60] used a deformable attention module to jointly solve the RS correction and deblurring problem. Unfortunately, they can only hallucinate one GS image at a specific moment, *e.g*. corresponding to the first [10] or central [24, 60] scanline time, and thus fall short of reconstructing a smooth and coherent GS video.

Very recently, Fan and Dai [9] developed the first rolling shutter temporal super-resolution network to extract a high framerate GS video from two consecutive RS images. It warps each RS frame to a latent GS frame corresponding to any of its scanlines through geometry-aware propagation. As a result, undesirable holes (*e.g*. black edges) appear due to the occlusion between the RS and GS images. Furthermore, it leverages a constant velocity motion assumption, which does not accurately capture the motion boundaries and produces artifacts around the moving objects. Two examples are shown in Figs. 1 and 6. In contrast, we propose a GS frame synthesis module, which is composed of contextual aggregation and motion enhancement layers, to reason about complex occlusions and motion patterns specific to moving objects, resulting in a significantly improved performance of GS video reconstruction.

3. RS-aware Frame Warping

RS image formation model. When an RS camera is in motion during the image acquisition, all its scanlines are exposed sequentially at different timestamps. Hence each scanline possesses a different local frame, as illustrated in Fig. 2. Without loss of generality, we assume that all pixels in the same row are exposed instantaneously at the same time. The number of rows in the image is h, and the constant inter-row delay time is τ_d. Therefore, the RS image formation model can be obtained as follows:

$$\lfloor \mathbf{I}^r(\mathbf{x}) \rfloor_s = \lfloor \mathbf{I}^g_s(\mathbf{x}) \rfloor_s , \tag{1}$$

where $\mathbf{I}^g_s$ is virtual GS images captured at time $\tau_d(s - h/2)$, and $\lfloor \cdot \rfloor_s$ denotes the extraction of pixel $\mathbf{x}$ in scanline s.

RS effect removal by forward warping. Since an RS image can be viewed as the result of successive row-by-row combinations of virtual GS image sequences within the imaging duration, one can invert the above RS imaging mechanism to remove RS distortions by

$$\mathbf{I}^r(\mathbf{x}) = \mathbf{I}^g_s(\mathbf{x} + \mathbf{u}_{r\to s}), \tag{2}$$

where $\mathbf{u}_{r\to s}$ is the displacement vector of pixel $\mathbf{x}$ from the RS image $\mathbf{I}^r$ to the virtual GS image $\mathbf{I}^g_s$. Stacking $\mathbf{u}_{r\to s}$ of all pixels yields a pixel-wise motion field, *a.k.a. undistortion flow* $\mathbf{U}_{r\to s}$, which can be used to RS-aware forward warping analogous to [9, 10, 24, 60]. However, when multiple pixels are mapped to the same location, forward warping is prone to suffer from conflicts, inevitably leading to overlapped pixels and holes. Softmax splatting [35] alleviates these problems by adaptively combining overlapping pixel information. Thus, the target GS frame corresponding to scanline s can be generated by

$$\hat{\mathbf{I}}^g_s = \mathcal{W}_F(\mathbf{I}^r, \mathbf{U}_{r\to s}), \tag{3}$$

where $\mathcal{W}_F$ represents the forward warping operator. We use softmax splatting in our implementation.

Problem setup. As depicted in Fig. 2, time t and scanline s correspond to each other. For compactness, in the following we will discard the symbol s and use the subscript t to denote the GS image $\mathbf{I}^g_t$ corresponding to time t. Following [12, 24, 62], we further assume that the readout time ratio [61], *i.e*. the ratio between the total scanline readout time (*i.e*. $h\tau_d$) and the inter-frame delay time, is equal to one. That is to say, the idle time between two adjacent RS frames is ignored in a short period of imaging time (*e.g*. < 50 ms). This is proved to be effective to account for the scanline-varying camera poses, avoiding non-trivial readout calibration [30]. Moreover, this also ensures temporally tractable frame interpolation for RS images. See the *supplementary material* for further instructions. Consequently, the central scanlines of the two consecutive RS images are recorded at time instances 0 and 1, respectively.

Given two RS frames $\mathbf{I}_0^r$ and $\mathbf{I}_1^r$ at adjacent times 0 and 1, we aim to synthesize an intermediate GS frame $\hat{\mathbf{I}}_t^g$, $t \in [0,1]$. This time interval is chosen because, as observed in [10], many details of the recovered GS images corresponding to time $t \in [-0.5,0) \cup (1,1.5]$ are more likely to be missing due to too much deviation from the temporal consistency.

3.1. Bilateral Motion Field Initialization

Network-based bilateral motion field (NBMF). To deliver each RS pixel $\mathbf{x}$ exposed at time τ (*i.e.* $\tau_0 \in [-0.5,0.5]$ or $\tau_1 \in [0.5,1.5]$, with subscripts indicating the image index) to the GS canvas corresponding to the camera pose at time $t \in [0,1]$, we need to estimate the motion field $\mathbf{U}_{0\to t}$ or $\mathbf{U}_{1\to t}$ (*cf*. Eq. (3)) to constrain each pixel's displacement. Note that the subscripts $0 \to t$ and $1 \to t$ indicate the RS-aware forward warping from RS images $\mathbf{I}_0^r$ and $\mathbf{I}_1^r$ to $\hat{\mathbf{I}}_t^g$, respectively. According to [9], we extend to the time dimension to model the BMF $\mathbf{U}_{0\to t}$ and $\mathbf{U}_{1\to t}$ by a scaling operation on the corresponding optical flow fields $\mathbf{F}_{0\to 1}$ and $\mathbf{F}_{1\to 0}$ between two consecutive RS frames, *i.e.*

$$\begin{aligned} \mathbf{U}_{0\to t}(\mathbf{x}) &= \mathbf{C}_{0\to t}(\mathbf{x}) \cdot \mathbf{F}_{0\to 1}(\mathbf{x}), \\ \mathbf{U}_{1\to t}(\mathbf{x}) &= \mathbf{C}_{1\to t}(\mathbf{x}) \cdot \mathbf{F}_{1\to 0}(\mathbf{x}), \end{aligned} \tag{4}$$

where

$$\begin{aligned} \mathbf{C}_{0\to t}(\mathbf{x}) &= \frac{(t-\tau_0)(h-\pi_v)}{h}, \\ \mathbf{C}_{1\to t}(\mathbf{x}) &= \frac{(\tau_1-t)(h+\pi_v')}{h}, \end{aligned} \tag{5}$$

represent the bilateral correction maps. π_v and π_v' encapsulate the underlying RS geometry [9] to reveal the inter-RS-frame vertical optical flow, depending on the camera parameters, the camera motion, and the depth and position of pixel $\mathbf{x}$. Furthermore, the BMF corresponding to different time steps t_1 and t_2 can be directly interconverted by

$$\mathbf{U}_{i\to t_2}(\mathbf{x}) = \frac{t_2-\tau}{t_1-\tau} \cdot \mathbf{U}_{i\to t_1}(\mathbf{x}), \quad i = 0,1. \tag{6}$$

Note that the motion field for RS removal has a significant time dependence (*a.k.a.* scanline dependence [9]). To capture the correction map in Eq. (5), a geometric optimization problem was posed in [61, 62] based on the differential formulation [11, 29]. Recently, as shown in Fig. 3 (a), an encoder-decoder network was proposed in [9] to essentially learn the underlying RS geometry, such that the BMF can be computed by Eq. (4) coupled with the estimated bidirectional optical flows, termed as NBMF. The arbitrary-time GS images are then generated by image warping based on explicit intra-frame propagation in Eq. (6). However, since the occlusion view is not available during warping, the resulting holes are visually unsatisfactory. Also, [9] is not adaptive to dynamic objects due to the reliance on a constant velocity motion assumption of the RS camera.

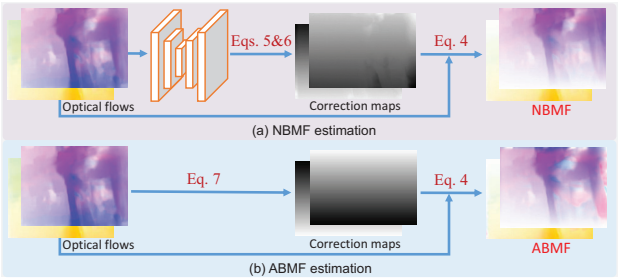

Figure 3. Illustration of the initial BMF estimation, including (a) NBMF and its approximation (b) ABMF.

Approximated bilateral motion field (ABMF). We observe that π_v and π_v' in Eq. (5) characterize the latent inter-GS-frame vertical optical flow, which are usually much smaller than the number of image rows h (*cf*. *supplementary materials* for in-depth analysis). Hence, we propose an approximated constraint $h - \pi_v \approx h \approx h + \pi_v'$ to rewrite Eq. (5) as:

$$\begin{aligned} \mathbf{C}_{0\to t}(\mathbf{x}) &= t - \tau_0, \\ \mathbf{C}_{1\to t}(\mathbf{x}) &= \tau_1 - t, \end{aligned} \tag{7}$$

where the time dependence is retained while the parallax effects (*i.e.* depth variation and camera motion) are neglected. That is, it is independent of the image content and can be pre-defined for a given image resolution. As depicted in Fig. 3 (b), such approximation is able to reach the correction map and then the ABMF via Eq. (4) in a simple and straightforward manner instead of relying on specialized deep neural networks. Note that the interconversion between varying ABMF satisfies Eq. (6) as well. The experimental results in Sec. 6.1 show that our ABMF, coupled with the contextual aggregation and motion enhancement, can serve as a strong and tractable baseline for GS frame synthesis.

4. Context-aware Video Reconstruction

We advocate recovering the intermediate global shutter image $\hat{\mathbf{I}}_t^g, t \in [0,1]$ from two input consecutive rolling shutter images $\mathbf{I}_0^r$ and $\mathbf{I}_1^r$. In this section, we will explain how to design a deep network to reason about time-aware motion profiles and occlusions, such that the photorealistic time-arbitrary GS image can be recovered faithfully.

4.1. Architecture Overview

As shown in Fig. 4, the proposed network consists of two modules, *i.e.* an NBMF-based or ABMF-based motion interpretation module, and a context (*i.e.* occlusions and partial dynamics) aware GS frame synthesis module. Firstly, we estimate the bidirectional optical flow fields $\mathbf{F}_{0\to 1}$ and $\mathbf{F}_{1\to 0}$ between $\mathbf{I}_0^r$ and $\mathbf{I}_1^r$, followed by the BMF estimation $\mathbf{U}_{0\to t}$ and $\mathbf{U}_{1\to t}$ via Eq. (4), which is based on NBMF (*i.e.* Eq. (5)) or ABMF (*i.e.* Eq. (7)), as illustrated in Fig. 3. Then, the input RS frames are forward warped using the initial bilateral motions, resulting in two initial intermediate GS frame candidates at time t. Finally, the GS frame

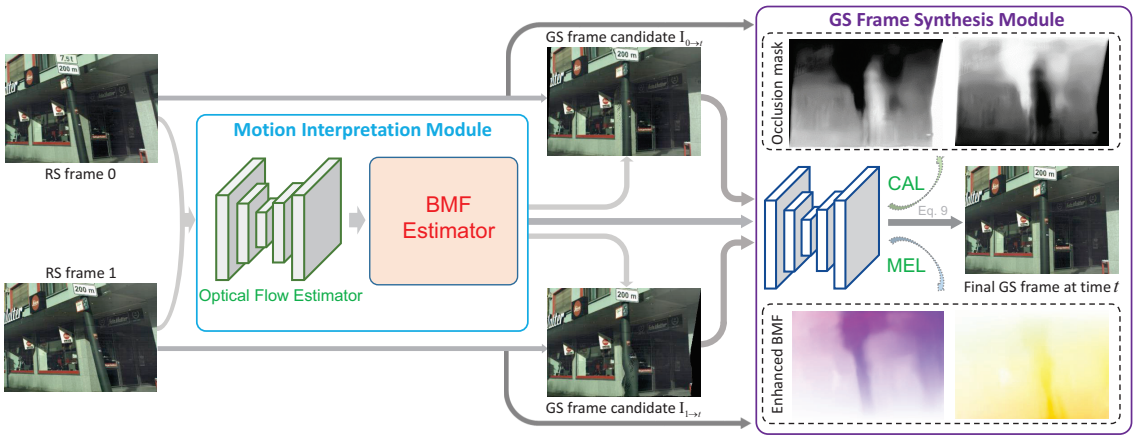

Figure 4. **Overall architecture.** It has two main processes. First, two initial GS frame candidates are obtained by the motion interpretation module. The details of BMF estimator (*i.e.* NBMF or ABMF) are elaborated in Fig. 3. Then, a GS frame synthesis module is proposed to reason about complex occlusions, motion profiles, and temporal abstractions to generate the final high-fidelity GS image at time $t \in [0, 1]$.

synthesis module takes the input RS frames, bidirectional optical flows, bilateral motion fields, and the initial intermediate GS frame candidates to synthesize the final GS reconstruction by aggregating the context information and compensating for the motion boundaries adaptively. Note that we empirically find that our ABMF-based CVR approach (called *CVR**) performs well despite its simplicity, while our NBMF-based CVR approach (called *CVR*) can further improve the quality of the final GS images.

Motion interpretation module $\mathcal{M}$ is composed of two submodules: an optical flow estimator and a bilateral motion field estimator. We first utilize the widely used PWC-Net [50] as the optical flow estimator to predict the bidirectional optical flow. To obtain an effective initial BMF, we follow [9] and use a dedicated encoder-decoder U-Net architecture [37,46], as shown in Fig. 3 (a), to estimate NBMF for forward warping, which is termed as $\mathcal{M}_N$. Particularly, $\mathcal{M}_N$ needs to be pre-trained by using the ground-truth (GT) central-scanline GS images for supervision. Alternatively, we propose to exploit its approximate version as shown in Fig. 3 (b), *i.e.* an ABMF-based motion interpretation module $\mathcal{M}_A$, to yield a simpler and faster prediction of the initial BMF. Finally, two initial intermediate GS frame candidates $\mathbf{I}^g_{0\to t}$ and $\mathbf{I}^g_{1\to t}$ can be generated by Eq. (3) based on the initial BMF estimations $\mathbf{U}_{0\to t}$ and $\mathbf{U}_{1\to t}$, respectively.

GS frame synthesis module $\mathcal{G}$ can be boiled down to two main layers: a motion enhancement layer (MEL) and a contextual aggregation layer (CAL). Note that some black holes and ambiguous misalignments may exist in the initial intermediate GS frame candidates due to heavy occlusions and partial moving objects, degrading the visual experience. Therefore, we aim at alleviating artifacts at the boundaries of dynamic objects and filling the occluded holes. Towards this goal, $\mathbf{I}^r_0$, $\mathbf{I}^r_1$, $\mathbf{F}_{0\to 1}$, $\mathbf{F}_{1\to 0}$, $\mathbf{U}_{0\to t}$, $\mathbf{U}_{1\to t}$, $\mathbf{I}^g_{0\to t}$, and $\mathbf{I}^g_{1\to t}$ are concatenated and fed into $\mathcal{G}$ to estimate the BMF residuals $\Delta\mathbf{U}_{0\to t}$ and $\Delta\mathbf{U}_{1\to t}$ and the bilateral occlusion masks $\mathbf{O}_{0\to t}$ and $\mathbf{O}_{1\to t}$. This time-aware occlusion mask is essential to guide GS frame synthesis to handle occlusions. We employ an encoder-decoder U-Net network [37, 46] as the backbone of $\mathcal{G}$, which has the same structure but different channels as the network in $\mathcal{M}_N$. The network is fully convolutional with skip connections and leaky ReLu activation functions. Besides, we leverage a sigmoid activation function on the output channels corresponding to the bilateral occlusion mask to limit its value between 0 and 1. Because $\mathcal{G}$ accepts cascades at different time instances, it can implicitly model the temporal abstraction to recover GS frames corresponding to arbitrary time step $t \in [0, 1]$.

Specifically, the final enhanced BMF can be obtained as:

$$\begin{aligned}\hat{\mathbf{U}}_{0\to t} &= \mathbf{U}_{0\to t} + \Delta\mathbf{U}_{0\to t},\\ \hat{\mathbf{U}}_{1\to t} &= \mathbf{U}_{1\to t} + \Delta\mathbf{U}_{1\to t},\end{aligned} \tag{8}$$

which can improve the quality of BMF by combining it with the proposed contextual consistency constraint, especially in motion boundaries and unsmooth regions. Subsequently, we can produce two refined intermediate GS frame candidates $\hat{\mathbf{I}}^g_{0\to t}$ and $\hat{\mathbf{I}}^g_{1\to t}$ by RS-aware forward warping in Eq. (3). Further, we assume that the content of the target GS image corresponding to $t \in [0, 1]$ can be recovered by at least one of the input RS images, which is promising as discussed in [10]. We therefore impose the constraint that $\mathbf{O}_{1\to t} = \mathbf{1} - \mathbf{O}_{0\to t}$. Intuitively, $\mathbf{O}_{0\to t}(\mathbf{x}) = 0$ implies $\mathbf{O}_{1\to t}(\mathbf{x}) = 1$, *i.e.* target pixels can be faithfully rendered by fully trusting $\mathbf{I}^r_1$, and vice versa. Similar to [18, 37, 56], we also take advantage of the temporal distances $1-t$ and t for the input RS frames $\mathbf{I}^r_0$ and $\mathbf{I}^r_1$, such that the temporally-closer pixels can be assigned a higher confidence. At last, the final intermediate GS frame $\hat{\mathbf{I}}^g_t$ can be synthesized by

$$\hat{\mathbf{I}}^g_t = \frac{(1-t)\mathbf{O}_{0\to t}\hat{\mathbf{I}}^g_{0\to t} + t\mathbf{O}_{1\to t}\hat{\mathbf{I}}^g_{1\to t}}{(1-t)\mathbf{O}_{0\to t} + t\mathbf{O}_{1\to t}}. \tag{9}$$

Input RS (Overlayed) | DiffSfM [61] | DiffHomo [62] | BMBC [38] | DAIN [2] | Cascaded method

DeepUnrollNet [24] | SUNet [10] | RSSR [9] | CVR* (Ours) | CVR (Ours) | Ground-truth

Figure 5. Qualitative results against baselines. Our method can successfully remove RS artifacts, yielding higher fidelity GS images.

Table 1. Quantitative comparisons on recovering GS images at time step $t = 0.5$. The numbers in **red** and blue represent the best and second-best performance. Our method is far superior to baseline methods and the proposed ABMF model is effective as an initialization.

Method	Runtime (seconds)	PSNR↑ (dB)			SSIM↑		LPIPS↓	
		CRM	CR	FR	CR	FR	CR	FR
DiffSfM [61]	467	24.20	21.28	20.14	0.775	0.701	0.1322	0.1789
DiffHomo [62]	424	19.60	18.94	18.68	0.606	0.609	0.1798	0.2229
DeepUnrollNet [24]	0.34	26.90	26.46	26.52	0.807	0.792	0.0703	0.1222
SUNet [10]	0.21	29.28	29.18	28.34	0.850	0.837	0.0658	0.1205
RSSR*	0.09	28.20	23.86	21.02	0.839	0.768	0.0764	0.1866
RSSR [9]	0.12	30.17	24.78	21.23	0.867	0.776	0.0695	0.1659
*CVR** (Ours)	0.12	31.82	31.60	28.62	0.927	0.845	0.0372	0.1117
CVR (Ours)	0.14	**32.02**	**31.74**	**28.72**	**0.929**	**0.847**	**0.0368**	**0.1107**

*: *applying our proposed approximated bilateral motion field (ABMF) model.*

4.2. Loss Function

Similar to [9, 24, 60], we use the reconstruction loss $\mathcal{L}_r$, the perceptual loss $\mathcal{L}_p$ [19], and the total variation loss $\mathcal{L}_{tv}$ to improve the quality of final GS and BMF predictions. Moreover, inspired by [10], we propose a contextual consistency constraint loss $\mathcal{L}_c$ to enforce the alignment of refined intermediate GS frame candidates with ground-truth, which is crucial to facilitate occlusion inference and motion compensation. In short, our loss function $\mathcal{L}$ is defined as:

$$\mathcal{L} = \lambda_r \mathcal{L}_r + \mathcal{L}_p + \lambda_c \mathcal{L}_c + \lambda_{tv} \mathcal{L}_{tv}, \tag{10}$$

where λ_r, λ_c and λ_{tv} are hyper-parameters. More details can be found in the *supplementary material*.

5. Experimental Setup

Datasets. We use the standard RS correction benchmark datasets [24] including Carla-RS and Fastec-RS, and divide the training and test sets as in [24]. The Carla-RS dataset is synthesized based on the Carla simulator [8], involving general 6-DOF camera motions. The Fastec-RS dataset records real-world RS images synthesized by a high-FPS GS camera mounted on a ground vehicle. Since they provide the first- and central-scanline GT supervisory signals, *i.e.* $t = 0$, 0.5, and 1, we utilize this triplet as GT to train our network. Note that we add a small perturbation to make Eq. (9) work properly, for example, transforming them to $t = 0.01$, 0.5, and 0.99, respectively. At the test phase, our method is capable of recovering GS video frames at any time $t \in [0, 1]$.

Training details. Our method is trained end-to-end using the Adam optimizer [21] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We empirically set $\lambda_r = 10$, $\lambda_c = 5$, and $\lambda_{tv} = 0.1$. The experiments are performed on an NVIDIA GeForce RTX 2080Ti GPU with a batch size of 4. We propose to train our network in two stages. Firstly, we solely train $\mathcal{M}$. To train the ABMF-based $\mathcal{M}_A$, we fine-tune PWC-Net [50] for 100 epochs from its pre-trained model on the RS benchmark in a self-supervised way [9, 20, 25, 54], and then ABMF can be computed directly and explicitly. Note that the training details of the NBMF-based $\mathcal{M}_N$ can be found in [9] with the supervision of central-scanline GT GS images. Secondly, we jointly train the entire model (*i.e.* $\mathcal{M}$ and $\mathcal{G}$) by $\mathcal{L}$ for another 50 epochs. At this time, the learning rate of $\mathcal{G}$ is set to 10^{-4} for training from scratch, and that of $\mathcal{M}$ is set to 10^{-5} for fine-tuning. We keep the vertical resolution constant and adopt a uniform random crop with a horizontal resolution of 256 pixels to augment the training data, similar to [9, 10] for better contextual exploration.

Evaluation strategies. As the Carla-RS dataset has the GT occlusion mask, we perform quantitative evaluations as follows: Carla-RS dataset with occlusion mask (*CRM*), Carla-RS dataset without occlusion mask (*CR*), and Fastec-RS dataset (*FR*). Standard metrics PSNR and SSIM, and learned perceptual metric LPIPS [58] are applied. Higher PSNR/SSIM or lower LPIPS score indicates better quality. Note that unless otherwise stated, *we refer to the GS images at time $t = 0.5$ for consistent comparisons.*

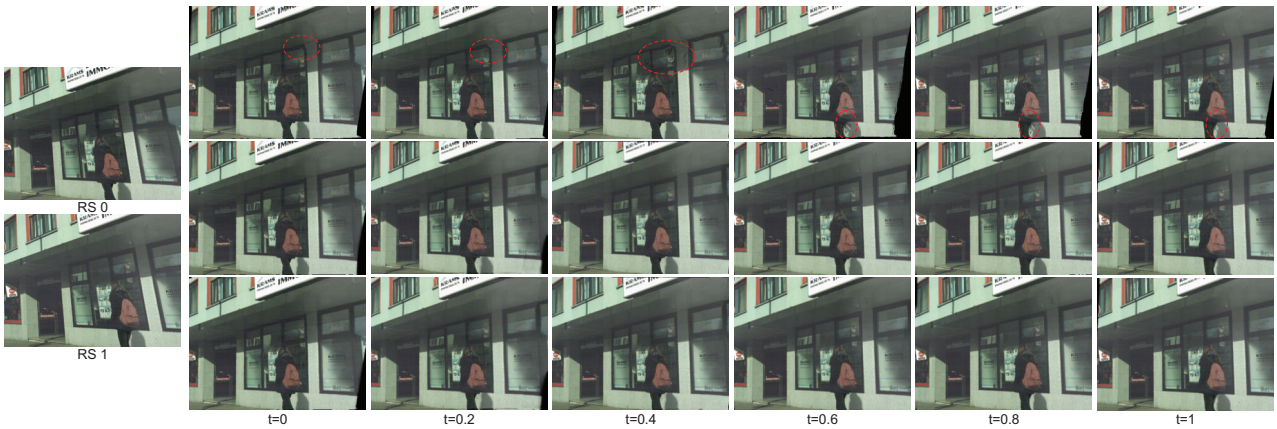

Figure 6. Example results of recovering six GS video images from the two input RS images (left column) by using RSSR [9], *CVR**, and *CVR* (three rows from top to bottom), respectively. Apart from many unfriendly black holes at the GS image edges, RSSR generates local errors and motion artifacts as shown in red circles. Our method can produce temporally consistent GS sequences with richer details.

Baselines. We perform comparisons with the following baselines. (i) **DiffSfM** [61] and **DiffHomo** [62] are traditional two-image based RS correction methods that require sophisticated optimization using RS models. (ii) **SUNet** [10] and **DeepUnrollNet** [24] recover only one GS frame from two consecutive RS frames by designing specialized CNNs. While **RSCD** [60] achieves this goal from three adjacent RS images. (iii) **RSSR** [9] generates a GS video from two consecutive RS images using deep learning, but suffers from black holes and motion artifacts. Moreover, we integrate the proposed ABMF model into RSSR to yield **RSSR***. (iv) **DAIN** [2] and **BMBC** [38] are SOTA VFI methods that are tailored for GS cameras. (v) **Cascaded method** generates two GS images sequentially from three consecutive RS inputs using DeepUnrollNet, and then interpolates in-between GS ones using DAIN. (vi) ***CVR*** and ***CVR**** are our proposed methods based on NBMF and ABMF, respectively. Note that our *RSSR**, RSSR, our *CVR**, and our *CVR* form a clear hierarchy of RS-based video reconstruction methods.

6. Results and Analysis

In this section, we compare with the baseline approaches and provide analysis and insight into our method.

6.1. Comparison with SOTA Methods

We report the quantitative and qualitative results in Table 1 and Fig. 5, respectively. Our proposed method achieves overwhelming dominance in RS effect removal, which is mainly attributed to context aggregation and motion pattern inference. Furthermore, although our proposed ABMF model is inferior to RSSR [9] when used to remove the RS effect (*i.e*. RSSR*), it can serve as a strong baseline for GS video frame reconstruction when combined with GS frame refinement. We believe that our hierarchical pipeline can provide a fresh perspective for the video reconstruction task with RS cameras. More results and analysis are shown in the *supplementary material*.

Note that our method is able to produce a continuous GS sequence, which is far beyond [10, 24, 60], although [10] can decode the plausible details of the GS image at a specific time. Traditional methods [61, 62] cannot estimate the underlying RS geometry robustly and accurately, resulting in ghosting artifacts. They are also computationally inefficient due to the complicated handling. Due to inherent flaws in the network architectures, the VFI methods [2, 38] fail to remove the RS effect. An intuitive cascade of RS correction and VFI methods tends to accumulate errors and is prone to blurring artifacts and local inaccuracies. Such cascades also have large models and thus be relatively time-consuming. In contrast, our end-to-end pipeline performs favorably against the SOTA methods in terms of both RS correction and inference efficiency. Note also that obnoxious black holes and object-specific motion artifacts appear in [9], degrading the visual experience, as outlined in Sec. 1. In general, our *CVR* improves RSSR and therefore recovers higher realism results, and our *CVR** also develops a new concise and efficient framework for related tasks.

6.2. GS Video Reconstruction Results

We apply our method to generate multiple in-between GS frames at arbitrary time $t \in [0, 1]$. The visual results for $5\times$ temporal upsampling are shown in Fig. 6. More results are provided in our *supplementary materials*. Our method can not only successfully remove the RS effect, but also can robustly reconstruct smooth and continuous GS videos.

6.3. Ablation Studies

Ablation on motion interpretation module $\mathcal{M}$. We first replace NBMF and ABMF with linear BMF (*i.e*. LBMF), which is a widely used BMF initialization scheme in popular VFI methods, *e.g*. [18, 34, 35, 38, 49]. Then, we replace PWC-Net with the SOTA optical flow estimation pipeline RAFT [51]. Finally, we freeze $\mathcal{M}$ and solely train $\mathcal{G}$ in the

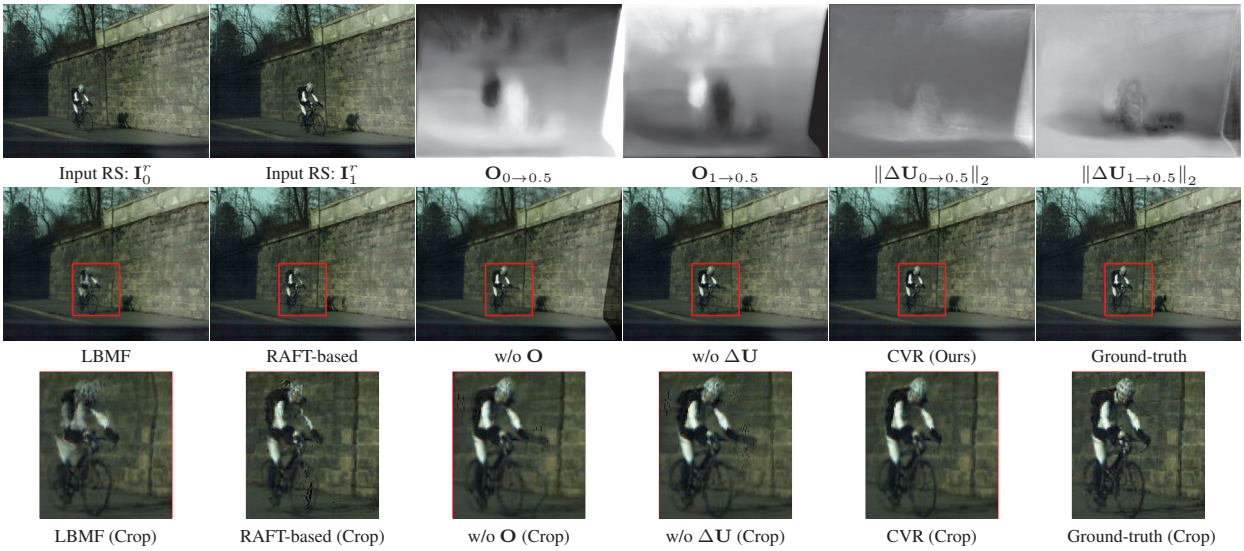

Figure 7. Visual results of ablation study. Our context-aware method is also adaptable to motion artifacts specific to moving objects.

Table 2. Ablation results for *CVR* architecture on $\mathcal{M}$, $\mathcal{G}$ and $\mathcal{L}$.

Settings	PSNR↑ (dB)			SSIM↑	
	CRM	CR	FR	CR	FR
LBMF	26.10	25.97	25.78	0.806	0.771
RAFT-based	30.50	29.89	27.99	0.917	0.840
Freeze $\mathcal{M}$	31.94	31.65	28.11	0.928	0.837
$\mathbf{T}\cdot\Delta\mathbf{U}$	32.00	31.63	28.56	**0.929**	0.845
w/o $\Delta\mathbf{U}$	31.90	31.65	28.32	0.928	0.841
w/o $\mathbf{O}$	28.22	26.31	24.04	0.902	0.813
w/o $\mathcal{L}_r$	31.80	31.53	28.31	0.927	0.840
w/o $\mathcal{L}_p$	31.60	31.34	28.49	**0.929**	0.842
w/o $\mathcal{L}_c$	31.88	31.64	28.44	0.928	0.842
w/o $\mathcal{L}_{tv}$	31.93	31.71	28.45	0.928	0.844
full model	**32.02**	**31.74**	**28.72**	**0.929**	**0.847**

training phase. As can be seen from Table 2 and Fig. 7, LBMF is extremely ineffective for the RS-based video construction task, which reveals the superiority of our proposed NBMF as well as ABMF. This could facilitate further research in related fields, especially the simpler ABMF. Since the RAFT-based full baseline is not easily optimized jointly end-to-end, it is prone to unsmoothness at local motion boundaries. Additionally, training the entire network together with $\mathcal{M}$ can improve model performance.

Ablation on GS frame synthesis module $\mathcal{G}$. We analyze the role of each component of $\mathcal{G}$ in Table 2, including **1**) multiplying $\Delta\mathbf{U}$ by a normalized scanline offset $\mathbf{T}$ to explicitly model its scanline dependence like [9,24,59], and **2**) removing MEL (*i.e*. w/o $\Delta\mathbf{U}$) and CAL (*i.e*. w/o $\mathbf{O}$), separately. Combined with Fig. 7, one can observe that they both lead to performance degradation, especially removing CAL, which causes aliasing effects during context aggregation, *e.g*. misaligned wheels and black edges. Moreover, removing MEL will reduce the adaptability of our method to object-specific motion artifacts, especially for the more challenging Fastec-RS dataset. In summary, our method can adaptively infer occlusions and enhance motion boundaries.

Ablation on loss function $\mathcal{L}$. We remove the loss terms one by one to analyze their respective roles. From Table 2, our loss function $\mathcal{L}$ is effective because it performs best when all loss terms are used.

6.4. Limitation and Discussion

Our method relies on optical flow estimation, so there may be aliasing artifacts in areas such as low/weak textures. Besides, although we have assumed that the pixels of the target GS image at time $t \in [0, 1]$ are visible in one of the RS images, some of them at the edges of the GS image may not be available, *e.g*. the lower right corner of GS images at $t = 0$ in Figs. 1 and 6, due to severe occlusions from fast camera motion or object motion. Future use of more frames may be able to fill in these possible invisible regions.

7. Conclusion

In this paper, we have presented a context-aware architecture *CVR* for end-to-end video reconstruction of RS cameras, which incorporates temporal smoothness to recover high-fidelity GS video frames with fewer artifacts and better details. Moreover, we have developed a simple yet efficient pipeline *CVR** based on the proposed ABMF model which works robustly with RS cameras. Our proposed framework exploits the spatio-temporal coherence embedded in the latent GS video via motion interpretation and occlusion reasoning, significantly outperforming the SOTA methods. We hope this study can shed light for future research on video frame reconstruction of RS cameras.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (61871325, 61901387), National Key Research and Development Program of China (2018AAA0102803), and Innovation Foundation for Doctor Dissertation of NWPU. The authors thank the anonymous reviewers for their valuable comments.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2505–2513, 2020. 3

[2] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019. 1, 2, 6, 7

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

[4] Zhixiang Chi, Rasoul Mohammadi Nasiri, Zheng Liu, Juwei Lu, Jin Tang, and Konstantinos N Plataniotis. All at once: temporally adaptive multi-frame interpolation with advanced motion modeling. In *Proceedings of European Conference on Computer Vision*, pages 107–123, 2020. 2

[5] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. Channel attention is all you need for video frame interpolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10663–10671, 2020. 2

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

[7] 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 IEEE International Conference on Computer Vision*, pages 2758–2766, 2015. 2

[8] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: an open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 6

[9] Bin Fan and Yuchao Dai. Inverting a rolling shutter camera: bring rolling shutter images to high framerate global shutter video. In *Proceedings of IEEE International Conference on Computer Vision*, pages 4228–4237, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[10] Bin Fan, Yuchao Dai, and Mingyi He. Sunet: symmetric undistortion network for rolling shutter correction. In *Proceedings of IEEE International Conference on Computer Vision*, pages 4541–4550, 2021. 1, 2, 3, 4, 5, 6, 7

[11] Bin Fan, Yuchao Dai, Zhiyuan Zhang, and Mingyi He. Fast and robust differential relative pose estimation with radial distortion. *IEEE Signal Processing Letters*, 29:294–298, 2021. 4

[12] Bin Fan, Ke Wang, Yuchao Dai, and Mingyi He. Rolling-shutter-stereo-aware motion estimation and image correction. *Computer Vision and Image Understanding*, 213:103296, 2021. 3

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

[14] Per-Erik Forssén and Erik Ringaby. Rectifying rolling shutter video from hand-held devices. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 507–514, 2010. 3

[15] Matthias Grundmann, Vivek Kwatra, Daniel Castro, and Irfan Essa. Calibration-free rolling shutter removal. In *Proceedings of IEEE International Conference on Computational Photography*, pages 1–8, 2012. 3

[16] Johan Hedborg, Per-Erik Forssén, Michael Felsberg, and Erik Ringaby. Rolling shutter bundle adjustment. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1434–1441, 2012. 1

[17] Sunghoon Im, Hyowon Ha, Gyeongmin Choe, Hae-Gon Jeon, Kyungdon Joo, and In So Kweon. Accurate 3d reconstruction from small motion clip for rolling shutter cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):775–787, 2018. 1

[18] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018. 1, 2, 5, 7

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of European Conference on Computer Vision*, pages 694–711, 2016. 6

[20] Rico Jonschkowski, Austin Stone, Jonathan T Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *Proceedings of European Conference on Computer Vision*, pages 557–572, 2020. 6

[21] Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015. 6

[22] Yizhen Lao and Omar Ait-Aider. A robust method for strong rolling shutter effects correction using lines with automatic feature selection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4795–4803, 2018. 3

[23] Yizhen Lao and Omar Ait-Aider. Rolling shutter homography and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2780–2793, 2021. 3

[24] Peidong Liu, Zhaopeng Cui, Viktor Larsson, and Marc Pollefeys. Deep shutter unrolling network. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5941–5949, 2020. 1, 2, 3, 6, 7, 8

[25] Peidong Liu, Joel Janai, Marc Pollefeys, Torsten Sattler, and Andreas Geiger. Self-supervised linear motion deblurring. *IEEE Robotics and Automation Letters*, 5(2):2475–2482, 2020. 6

[26] Yihao Liu, Liangbin Xie, Li Siyao, Wenxiu Sun, Yu Qiao, and Chao Dong. Enhanced quadratic video interpolation. In *Proceedings of European Conference on Computer Vision*, pages 41–56, 2020. 2

[27] Yu-Lun Liu, Yi-Tung Liao, Yen-Yu Lin, and Yung-Yu Chuang. Deep video frame interpolation using cyclic frame generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8794–8802, 2019. 2

[28] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of IEEE International Conference on Computer Vision*, pages 4463–4471, 2017. 2

[29] Yi Ma, Jana Košecká, and Shankar Sastry. Linear differential algorithm for motion recovery: a geometric approach. *International Journal of Computer Vision*, 36(1):71–89, 2000. 4

[30] Marci Meingast, Christopher Geyer, and Shankar Sastry. Geometric models of rolling-shutter cameras. *arXiv preprint arXiv:cs/0503076*, 2005. 3

[31] Simone Meyer, Abdelaziz Djelouah, Brian McWilliams, Alexander Sorkine-Hornung, Markus Gross, and Christopher Schroers. Phasenet for video frame interpolation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 498–507, 2018. 2

[32] Simone Meyer, Oliver Wang, Henning Zimmer, Max Grosse, and Alexander Sorkine-Hornung. Phase-based frame interpolation for video. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1410–1418, 2015. 2

[33] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. Ccvs: context-aware controllable video synthesis. In *Proceedings of Advances in Neural Information Processing Systems*, volume 34, 2021. 2

[34] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1701–1710, 2018. 2, 7

[35] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5437–5446, 2020. 2, 3, 7

[36] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 670–679, 2017. 2

[37] Avinash Paliwal and Nima Khademi Kalantari. Deep slow motion video reconstruction with hybrid imaging system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1557–1569, 2020. 5

[38] Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. Bmbc: bilateral motion estimation with bilateral cost volume for video interpolation. In *Proceedings of European Conference on Computer Vision*, pages 109–125, 2020. 1, 2, 6, 7

[39] Junheum Park, Chul Lee, and Chang-Su Kim. Asymmetric bilateral motion estimation for video frame interpolation. In *Proceedings of IEEE International Conference on Computer Vision*, pages 14539–14548, 2021. 1, 2

[40] Pulak Purkait and Christopher Zach. Minimal solvers for monocular rolling shutter compensation under ackermann motion. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, pages 903–911, 2018. 3

[41] Pulak Purkait, Christopher Zach, and Ales Leonardis. Rolling shutter correction in Manhattan world. In *Proceedings of IEEE International Conference on Computer Vision*, pages 882–890, 2017. 3

[42] Fitsum A Reda, Deqing Sun, Aysegul Dundar, Mohammad Shoeybi, Guilin Liu, Kevin J Shih, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Unsupervised video interpolation using cycle consistency. In *Proceedings of IEEE International Conference on Computer Vision*, pages 892–900, 2019. 2

[43] Vijay Rengarajan, Yogesh Balaji, and AN Rajagopalan. Unrolling the shutter: cnn to correct motion distortions. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2291–2299, 2017. 1, 3

[44] Vijay Rengarajan, Ambasamudram N Rajagopalan, and Rangarajan Aravind. From bows to arrows: rolling shutter rectification of urban scenes. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2773–2781, 2016. 2, 3

[45] Erik Ringaby and Per-Erik Forssén. Efficient video rectification and stabilisation for cell-phones. *International Journal of Computer Vision*, 96(3):335–352, 2012. 3

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241, 2015. 5

[47] Olivier Saurer, Kevin Koser, Jean-Yves Bouguet, and Marc Pollefeys. Rolling shutter stereo. In *Proceedings of IEEE International Conference on Computer Vision*, pages 465–472, 2013. 3

[48] David Schubert, Nikolaus Demmel, Lukas von Stumberg, Vladyslav Usenko, and Daniel Cremers. Rolling-shutter modelling for direct visual-inertial odometry. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2462–2469, 2019. 1

[49] Li Siyao, Shiyu Zhao, Weijiang Yu, Wenxiu Sun, Dimitris Metaxas, Chen Change Loy, and Ziwei Liu. Deep animation video interpolation in the wild. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6587–6595, 2021. 2, 7

[50] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 2, 5, 6

[51] Zachary Teed and Jia Deng. Raft: recurrent all-pairs field transforms for optical flow. In *Proceedings of European Conference on Computer Vision*, pages 402–419, 2020. 2, 7

[52] Subeesh Vasu, Mahesh MR Mohan, and AN Rajagopalan. Occlusion-aware rolling shutter rectification of 3d scenes. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 636–645, 2018. 1

[53] Ke Wang, Bin Fan, and Yuchao Dai. Relative pose estimation for stereo rolling shutter cameras. In *Proceedings of IEEE International Conference on Image Processing*, pages 463–467, 2020. 3

[54] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion aware unsupervised learning

of optical flow. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4884–4893, 2018. 6

[55] Huicong Wu, Liang Xiao, and Zhihui Wei. Simultaneous video stabilization and rolling shutter removal. *IEEE Transactions on Image Processing*, 30:4637–4652, 2021. 1

[56] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In *Proceedings of Advances in Neural Information Processing Systems*, volume 32, 2019. 1, 2, 5

[57] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 2

[58] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[59] Zhihang Zhong, Mingdeng Cao, Xiao Sun, Zhirong Wu, Zhongyi Zhou, Yinqiang Zheng, Stephen Lin, and Imari Sato. Bringing rolling shutter images alive with dual reversed distortion. *arXiv preprint arXiv:2203.06451*, 2022. 8

[60] Zhihang Zhong, Yinqiang Zheng, and Imari Sato. Towards rolling shutter correction and deblurring in dynamic scenes. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9219–9228, 2021. 2, 3, 6, 7

[61] Bingbing Zhuang, Loong-Fah Cheong, and Gim Hee Lee. Rolling-shutter-aware differential sfm and image rectification. In *Proceedings of IEEE International Conference on Computer Vision*, pages 948–956, 2017. 1, 2, 3, 4, 6, 7

[62] Bingbing Zhuang and Quoc-Huy Tran. Image stitching and rectification for hand-held cameras. In *Proceedings of European Conference on Computer Vision*, pages 243–260, 2020. 1, 2, 3, 4, 6, 7

[63] Bingbing Zhuang, Quoc-Huy Tran, Pan Ji, Loong-Fah Cheong, and Manmohan Chandraker. Learning structure-and-motion-aware rolling shutter correction. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4551–4560, 2019. 1, 2, 3