Self-Supervised Learning for Rolling Shutter Temporal Super-Resolution

Bin Fan, Ying Guo, Yuchao Dai, *Member, IEEE*, Chao Xu, and Boxin Shi, *Senior Member, IEEE*

***Abstract*—Most cameras on portable devices adopt a rolling shutter (RS) mechanism, encoding sufficient temporal dynamic information through sequential readouts. This advantage can be exploited to recover a temporal sequence of latent global shutter (GS) images. Existing methods rely on fully supervised learning, necessitating specialized optical devices to collect paired RS-GS images as ground-truth, which is too costly to scale. In this paper, we propose a self-supervised learning framework for the first time to produce a high frame rate GS video from two consecutive RS images, unleashing the potential of RS cameras. Specifically, we first develop the unified warping model of RS2GS and GS2RS, enabling the complement conversions of RS2GS and GS2RS to be incorporated into a uniform network model. Then, based on the cycle consistency constraint, given a triplet of consecutive RS frames, we minimize the discrepancy between the input middle RS frame and its cycle reconstruction, generated by interpolating back from the predicted two intermediate GS frames. Experiments on various benchmarks show that our approach achieves comparable or better performance than state-of-the-art supervised methods while enjoying stronger generalization capabilities. Moreover, our approach makes it possible to recover smooth and distortion-free videos from two adjacent RS frames in the real-world BS-RSC dataset, surpassing prior limitations.**

***Index Terms*—Rolling shutter, Temporal super-resolution, Self-supervised learning, Cycle consistency.**

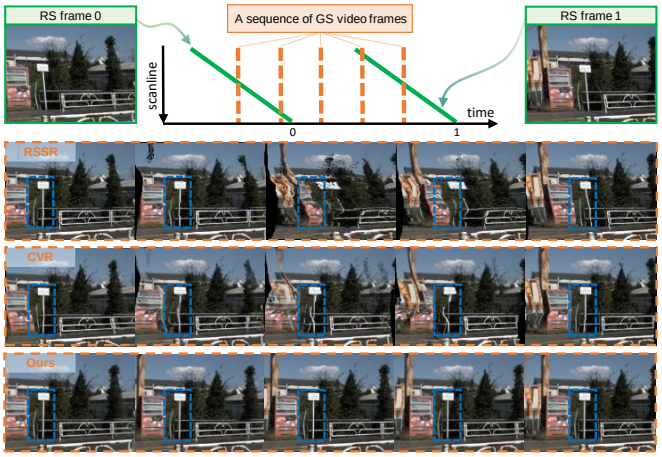

Fig. 1. An RS temporal super-resolution example on the real-world BS-RSC dataset [7]. We propose the first self-supervised learning method to recover a sequence of GS video frames from two consecutive RS frames. Due to the lack of sufficient paired RS-GS supervision signals, state-of-the-art supervised methods (*e.g.*, RSSR [8] and CVR [9]) fail to synthesize the temporal sequences of latent GS images in the RGB color space. By contrast, benefiting from the proposed self-supervised learning framework, our approach can accurately produce a high frame-rate GS video with richer details and better temporal consistency, *e.g.*, the signpost in the blue box.

I. Introduction

CMOS image sensors are the mainstream choice for mobile phones and low-budget commercial cameras due to their low cost and simple manufactures [1], [2]. Most electronic CMOS cameras employ a rolling shutter (RS) mechanism in which pixels are exposed in a row-wise fashion. Unlike global shutter (GS) cameras, which instantly take a snapshot of the entire scene, the inter-row readout delay of RS cameras causes distracting motion distortions known as the RS effect. For instance, straight lines are skewed and the image content appears to wobble, which is increasingly becoming a common nuisance in photography [3]–[6].

To improve the practical applicability of RS cameras, an intuitive motivation is to remove the RS effect, thus spawning a spectrum of RS correction methods [10]–[15]. Existing methods advocate recovering a single time-specific GS image from adjacent RS frames, *e.g.*, corresponding to the first [11], [16], [17] or middle [7], [10], [18], [19] scanline time. As such, they essentially do not alter the frame rate of the input video. Considering the excellent temporal dynamics of RS cameras, one would expect to not only eliminate the RS effect but also to revive and relive all time-arbitrary latent views as beheld by a virtual GS camera. Such a more challenging expectation has the potential to go beyond the temporal resolution limit of commercial cameras. Unfortunately, despite the remarkable success of video frame interpolation (VFI) methods [20]–[22], they cannot be applied to RS cameras to generate satisfactory intermediate GS frames, leading to residual RS distortions [9]. This is because the VFI method inherently formulates frame warping in a GS-adaptive manner, making it suitable solely for GS-image-to-GS-video (GS2GS) conversion, but not for RS-image-to-GS-video (RS2GS) conversion.

Towards the goal of RS2GS conversion, RS temporal super-resolution (RSTSR) [8] emerges as an effective solution that jointly handles VFI and RS correction, *i.e.*, interpolating a crisp and pleasing GS video to bring RS images alive. The most crucial step in this process is the construction of the RS2GS warping model. Typically, constant velocity [8], [9] and constant acceleration [23], [24] camera motion assumptions are

Bin Fan and Chao Xu are with the National Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University, Beijing 100871, China (e-mail: binfan@pku.edu.cn, xuchao@cis.pku.edu.cn).

Ying Guo and Yuchao Dai are with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, Shaanxi 710129, China (e-mail: yingguo@mail.nwpu.edu.cn; daiyuchao@nwpu.edu.cn).

Boxin Shi is with the State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing 100871, China (e-mail: shiboxin@pku.edu.cn).

Boxin Shi is the corresponding author.

TABLE I
OVERVIEW OF UNROLLING THE SHUTTER FROM TWO ADJACENT FRAMES.

Method	Time-Specific				Time-Arbitrary		
	DSUN [10]	SUNet [11]	CIEUnroll [16]	JAMNet [19]	RSSR [8]	CVR [9]	SelfRSSplat (Ours)
Feature-based synthesis	✓	✓	✓	✓	✗	✗	✓
RS2GS conversion	✗	✗	✗	✗	✓	✓	✓
GS2RS conversion	✗	✗	✗	✗	✗	✗	✓
Self-supervision	✗	✗	✗	✗	✗	✗	✓

harnessed to model the underlying spatio-temporal coherence, so that the undistortion flow can be estimated for frame warping through RS2GS modeling. Despite the promising results, as outlined in Table I, state-of-the-art RSTSR methods, *e.g.*, RSSR [8], CVR [9], hinge on fully supervised training and are only tailored for RS2GS conversion, resulting in two obvious limitations:

- **Expensive GT acquisition**. There are two main means to collect ground-truth (GT) RS-GS image pairs. On the one hand, the RS image can be simulated [10] by row-by-row splicing from high frame rate GS videos, but this suffers from stripe artifacts and domain gaps with real data. On the other hand, a well-designed beam-splitter system [7], [18] can be used in practice, but this requires rigorous optical registration and time synchronization. These factors lead to costly GT acquisition of paired GS images and also affect the algorithm's generalization to new RS data (see Fig. 1) where GS GT is limited or unavailable.
- **Insufficient model scalability**. Current RSTSR methods, although constructing the RS2GS warping model, are not yet adaptable to GS-image-to-RS-video (GS2RS) conversion, which is a pivotal element in achieving self-supervision learning. For GS2RS conversion, recent efforts either leverage the pre-trained VFI method to synthesize a high frame rate GS video sequence, followed by row-by-row stitching [16], [25], [26], or engineer a complex and independent warping function for dual-reversed RS setups [27]. However, none of these strategies can assign the two complement conversion attributes (*i.e.*, RS2GS and GS2RS conversions) to a common network model, which calls for an urgent need for a unified model representation of RS2GS and GS2RS.

To address the above limitations, we propose a *self-supervised* RS temporal super-resolution method to unlock the potential of RS cameras fully. It is trained in a self-supervised cycle consistent fashion, canceling the necessity of paired GT RS-GS images and shrinking the domain gap between synthetic and real data. To the best of our knowledge, this is the first attempt at the self-supervised RSTSR task. The key idea is inspired by [28], [29], which is based on the cycle consistency, where triplet samples with consecutive frames are constructed to exploit inter-frame consistency. Furthermore, we derive a *unified* model for RS2GS and GS2RS, which is parameter-free and flexible, allowing the two desired conversion attributes to be incorporated into a uniform network model simultaneously.

Specifically, we first inject the unified model of RS2GS and GS2RS into the motion estimation module of the VFI method (*i.e.*, SoftSplat [21]) to conceptually adapt both RS2GS and GS2RS conversions, termed RSSplat. It forward warps the RS (or GS) context features based on the bidirectional motion fields estimated by the RS2GS (or GS2RS) formulation, and then decodes the target GS (or RS) image in the feature space. We found that this feature-based synthesis architecture along the lines of [10], [11], [16], [19] has superior performance than CVR [9] that performs frame synthesis in RGB space, which is why we base our self-supervised approach on RSSplat instead of CVR (see Sec. IV-B). Then, in the context of cycle consistency, we optimize RSSplat to minimize the difference between the input second RS frame and its cycle reconstruction, obtained by interpolating back from the estimated two latent GS frames. In this way, we equip RSSplat with the self-supervised learning ability, termed **SelfRSSplat**[1], to generalize to RS videos with limited or no GS GT references. Meanwhile, a simple yet effective color consistency loss is imposed to combat color degradation. Extensive experimental results demonstrate that our SelfRSSplat achieves competitive performance with prior supervised methods while exhibiting stronger generalization. Interestingly, our SelfRSSplat outperforms state-of-the-art (SoTA) methods on the LPIPS metric.

In a nutshell, our main contributions can be summarized as:

- We propose for the first time a self-supervised learning framework for temporal super-resolution with RS images by applying cycle consistency.
- We develop a unified model for RS2GS and GS2RS, which endows a common network model with both complement conversion attributes of RS2GS and GS2RS, ensuring self-supervised training.
- Our approach is capable of unrolling two adjacent RS images from the *GT-limited* real-world BS-RSC dataset into a smooth and coherent GS video, as illustrated in Fig. 1, unlocking the potential of RS cameras.

II. RELATED WORK

Video frame interpolation. VFI is a long-standing computer vision research topic, where deep learning-based approaches have shown extraordinary performance in the recent literature [30]–[33]. Depending on whether optical flow is utilized, it can be roughly divided into two categories: flow-based [20], [21], [34]–[36] and flow-agnostic [37]–[40]. With the rapid development of optical flow estimation techniques [41]–[43], flow-based VFI methods have achieved dominance. Notably, linear [20], [21], [44], quadratic [45], quasi-quadratic [46], cubic [47], and hybrid [48], [49] motion interpolation schemes were developed. In addition, efficient frame synthesis architectures were designed, including contextual warping [50], [51], occlusion inference [52], [53], and transformer [54], [55], *etc.* Yet, all these VFI methods assume that the camera adopts a GS mechanism, which hinders the generation of satisfactory intermediate GS frames in the case of input RS images. In this paper, we propose a tractable temporal super-resolution method specifically for RS cameras, which can effectively remove RS artifacts while synthesizing in-between frames.

RS temporal super-resolution. Recently, extracting latent high frame rate GS videos hidden in RS images has received

[1]Code is available at https://github.com/GitCVfb/SelfRSSplat.

increasing attention [6]. The pioneer work directly warped the RS image based on geometric propagation models driven by constant velocity [8] or constant acceleration [23] to generate time-arbitrary GS images. Subsequently, Fan *et al.* [9] presented a context-aware video reconstruction network, CVR, to perform occlusion reasoning and motion compensation, resulting in impressive results. In addition to the above exploration of spatio-temporal coherence from a minimal configuration of two adjacent RS frames, Qu *et al.* [24] proposed a quadratic motion model to handle non-linear movements and complex occlusions via five consecutive RS inputs. Moreover, additional global reset features [56], [57], and dual-reversed RS images [25], [58] were also fully exploited to better invert the RS imaging mechanism. Unfortunately, all these methods require supervised training based on datasets containing paired GT RS-GS images, which poses huge challenges to both GS GT collection and model deployment in real-world scenarios. In contrast, we propose for the first time a self-supervised RSTSR method, as detailed in Table I, which can be trained in a cycle-consistent manner, eliminating the necessity of GS GT references.

Although Shang *et al.* [27] performed self-supervised video reconstruction from dual-reversed RS images by applying cycle consistency, it is inherently different from our method. In the problem setting, the input of [27] is dual-reversed RS images, while ours is two consecutive RS images. They have different RS geometries and thus essentially belong to two different research areas. Compared with consecutive RS input, the dual-reversed RS image records an extra row of pixels during each exposure and contains a shorter time step (*i.e.*, single frame imaging time) with smaller pixel displacement. As a result, the dual-reversed RS setup is more friendly for temporal super-resolution, as pointed out in [25]. Thus, cyclic consistency can be fulfilled in [27] by directly warping the estimated GS images to the input RS domain without introducing additional RS frames. However, this strategy would become infeasible for consecutive RS setups due to longer time steps. For two consecutive RS inputs, both RS2GS and GS2RS need to be performed in a common network to ensure cycle consistency, which is challenging yet indispensable. We develop a unified model of RS2GS and GS2RS and introduce an additional RS frame for self-supervised training.

Cycle consistency. As a tractable constraint, it has been widely applied in various self-supervised methods, such as correspondence learning [59], [60], visual representation [61]–[63], and image-to-image translation [64]. When challenged with the video interpolation task, its core idea is to reconstruct back frames from the original temporal resolution by interpolating between predicted intermediate frames [28], [29], [65], [66]. The method closest to our work is UnSuperSloMo [28], which optimizes the seminal VFI model (*i.e.*, SuperSloMo [20]) based on a given triplet of consecutive frames to minimize the discrepancy between the original middle frame and its cycle reconstruction. However, the resulting model is only capable of GS2GS and cannot yet be used for both RS2GS and GS2RS. Note that RS2GS and GS2RS are the two essential complement elements to achieve self-supervised RSTSR under the cycle-consistency criterion, which poses significant challenges for model design. In this paper, we build the analytical model of RS2GS and GS2RS under a unified framework, which allows RS2GS and GS2RS conversions with shared parameters, thereby supporting self-supervised cycle consistency training.

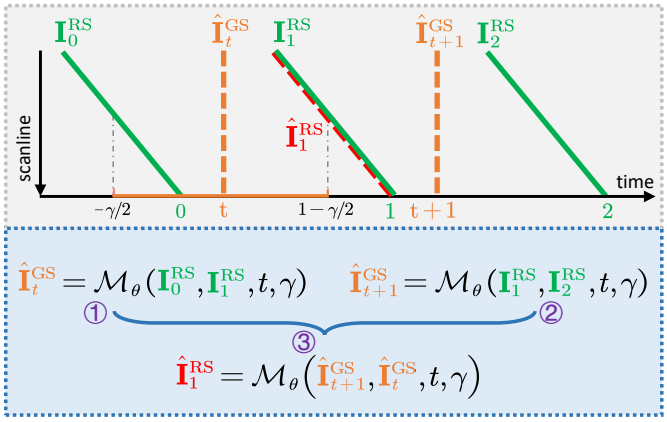

Fig. 2. Schematic of the cycle consistency for self-supervised RS temporal super-resolution. We first ① ② synthesize two intermediate GS images (*i.e.*, $\hat{\mathbf{I}}_t^{GS}, \hat{\mathbf{I}}_{t+1}^{GS}$) between each two input RS images (*i.e.*, $\{\mathbf{I}_0^{RS}, \mathbf{I}_1^{RS}\}$ and $\{\mathbf{I}_1^{RS}, \mathbf{I}_2^{RS}\}$), and then ③ reuse them to predict the original second RS image as $\hat{\mathbf{I}}_1^{RS}$. Finally, the difference can be computed between $\mathbf{I}_1^{RS}$ and its cycle reconstruction $\hat{\mathbf{I}}_1^{RS}$, which fully leverages the consistency within video sequences. Note that these three steps, whether ① ② RS2GS or ③ GS2RS conversion, adopt a uniform network model $\mathcal{M}$ with learnable parameters θ.

III. Method

In this section, we first provide a mathematical definition of the RSTSR task in the general RS scenario in Sec. III-A, and introduce how the self-supervised learning framework is designed under the cycle consistency constraint in Sec. III-B. Then, we develop a unified formulation for RS2GS and GS2RS in Sec. III-C, laying the foundation for arranging the two complement conversion attributes within a uniform network model to achieve self-supervised training. Finally, we briefly display the architecture of the proposed self-supervised RSTSR method in Sec. III-D and summarize the loss function in Sec. III-E.

A. Definition of the RS Temporal Super-resolution Task

Given two consecutive RS images $\mathbf{I}_0^{RS}$ and $\mathbf{I}_1^{RS}$, we define the RSTSR task similarly to CVR [9], aiming to restore a temporal sequence of latent GS images, whose start and end timestamps correspond to the middle scanline times of $\mathbf{I}_0^{RS}$ and $\mathbf{I}_1^{RS}$, respectively. Assuming an RS camera with a readout time ratio of γ, where $0 < \gamma \leq 1$ represents the ratio between the total readout time and the total imaging time of an RS image frame. As an RS-specific parameter, γ can be calibrated by [67], [68] and is widely used in RS geometric modeling [8], [9], [17], [69], [70]. Without loss of generality, in this paper, we denote the last scanline times of $\mathbf{I}_0^{RS}$ and $\mathbf{I}_1^{RS}$ as 0 and 1, respectively, which facilitates the unification of the two complement conversions (*i.e.*, RS2GS and GS2RS) later. Further, it is easy to determine that the middle scanline times of $\mathbf{I}_0^{RS}$ and $\mathbf{I}_1^{RS}$ are $-\gamma/2$ and $1-\gamma/2$, respectively.

Therefore, as shown by the solid orange line segment at the top of Fig. 2, the RSTSR task can be defined more

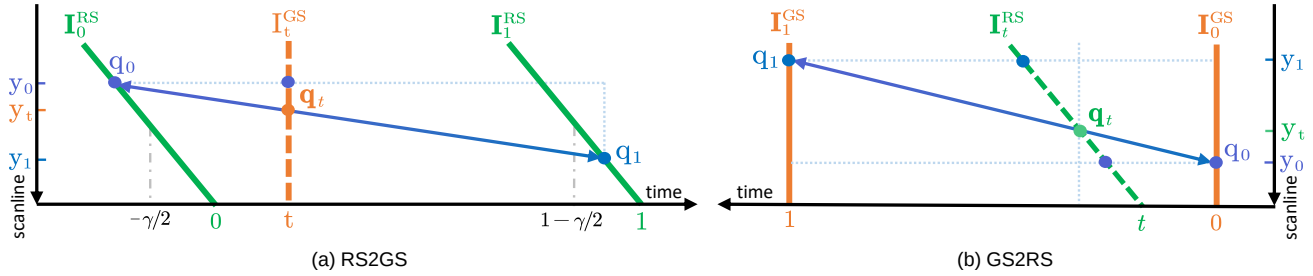

Fig. 3. Illustration of unified formulation of RS2GS and GS2RS, aiming to approximate bidirectional motion fields for frame warping.

generally in the time dimension as recovering the time-arbitrary intermediate GS image $\hat{\mathbf{I}}_t^{\mathrm{GS}}$ at time $t \in [-\gamma/2, 1-\gamma/2]$ from two neighboring RS images $\mathbf{I}_0^{\mathrm{RS}}$ and $\mathbf{I}_1^{\mathrm{RS}}$. Formally,

$$\hat{\mathbf{I}}_t^{\mathrm{GS}} = \mathcal{M}_\theta(\mathbf{I}_0^{\mathrm{RS}}, \mathbf{I}_1^{\mathrm{RS}}, t, \gamma), \tag{1}$$

where $\mathcal{M}$ denotes an RSTSR network model with learnable parameters θ. Note that the off-the-shelf RSTSR methods, *e.g.*, RSSR [8], CVR [9], require fully supervised training to learn a feasible θ, which hinders video reconstruction in real-world scenarios where GT RS-GS image pairs are limited or even lacking. Below, we aim to answer the core question: *Given only two consecutive RS images without corresponding GT RS-GS image pairs, how to design a self-supervised learning framework to learn θ effectively?*

B. Overview of Self-Supervised Learning

Inspired by the cycle consistency constraint, which is first proposed in CycleGAN [64] and then becomes a tractable constraint in self-supervised VFI methods [28], [29], [65], we design a self-supervised learning framework for RSTSR. The core design philosophy is that for a given triplet of input RS image sequence, if we generate an intermediate GS frame between each two consecutive RS frames, and generate back their intermediate RS frame, the resulting RS frame must match the original middle RS frame. Specifically, as illustrated in Fig. 2, given a triplet of consecutive RS frames $\{\mathbf{I}_0^{\mathrm{RS}}, \mathbf{I}_1^{\mathrm{RS}}, \mathbf{I}_2^{\mathrm{RS}}\}$, we first synthesize an in-between GS frame $\hat{\mathbf{I}}_t^{\mathrm{GS}}$ from $\mathbf{I}_0^{\mathrm{RS}}$ and $\mathbf{I}_1^{\mathrm{RS}}$ according to Eq. (1), and then similarly, the latent GS frame $\hat{\mathbf{I}}_{t+1}^{\mathrm{GS}}$ corresponding to time $t+1$ can be yielded from $\mathbf{I}_1^{\mathrm{RS}}$ and $\mathbf{I}_2^{\mathrm{RS}}$, *i.e.*,

$$\hat{\mathbf{I}}_{t+1}^{\mathrm{GS}} = \mathcal{M}_\theta(\mathbf{I}_1^{\mathrm{RS}}, \mathbf{I}_2^{\mathrm{RS}}, t, \gamma). \tag{2}$$

Finally, based on these two predicted intermediate GS proposals $\{\hat{\mathbf{I}}_t^{\mathrm{GS}}, \hat{\mathbf{I}}_{t+1}^{\mathrm{GS}}\}$, we can reconstruct back the input second RS frame as follows:

$$\hat{\mathbf{I}}_1^{\mathrm{RS}} = \mathcal{M}_\theta(\hat{\mathbf{I}}_{t+1}^{\mathrm{GS}}, \hat{\mathbf{I}}_t^{\mathrm{GS}}, t, \gamma). \tag{3}$$

As a result, the model parameters θ can be updated by minimizing the difference between the input RS image $\mathbf{I}_1^{\mathrm{RS}}$ and its cycle reconstruction $\hat{\mathbf{I}}_1^{\mathrm{RS}}$, *i.e.*,

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\mathbf{I}_1^{\mathrm{RS}}, \hat{\mathbf{I}}_1^{\mathrm{RS}}), \tag{4}$$

where η is the adaptive learning rate and $\mathcal{L}$ represents the loss function. Such a cycle consistency framework enables the model to make full use of its own inter-frame consistency properties in a self-supervised fashion via Eqs. (1), (2), (3), thereby obtaining an effective θ to bring RS images alive. Note that this workflow contains *two key elements*: One is the need to jointly handle two complement view conversions (*i.e.*, RS2GS in Eqs. (1), (2), and GS2RS in Eq. (3)); the other is the network model $\mathcal{M}$ used in these conversions should have the same parameters θ. Next, we integrate the two complement attributes into a uniform network through RS2GS and GS2RS formulations.

Tackling degenerate solutions. A degenerate solution to the optimization process of Eq. (4) might be to copy the input RS frame $\mathbf{I}_1^{\mathrm{RS}}$ as the intermediate prediction. However, this does not occur in our learning setting because we utilize diverse t in Eqs. (1) and (2) to estimate intermediate GS frames and $\mathbf{I}_1^{\mathrm{RS}}$ is located at different input positions, thus avoiding trivial solutions. Note that another possible degradation is that the colors of the intermediate GS images $\hat{\mathbf{I}}_t^{\mathrm{GS}}$ and $\hat{\mathbf{I}}_{t+1}^{\mathrm{GS}}$ may be distorted, as displayed in Fig. 4. While this does not affect the final prediction of $\hat{\mathbf{I}}_1^{\mathrm{RS}}$ during cycle consistency training, it leads to visually unpleasant $\hat{\mathbf{I}}_t^{\mathrm{GS}}$ when inferring via Eq. (1). To this end, we propose a color consistency loss in Sec. III-E that forces the warped GS proposal to be consistent with the input RS image in RGB space.

Fig. 4. Degenerate solutions during training. Without imposing the color consistency loss, the intermediate GS predictions $\hat{\mathbf{I}}_t^{\mathrm{GS}}$ (the desired output) and $\hat{\mathbf{I}}_{t+1}^{\mathrm{GS}}$ may suffer from color distortions.

C. Unified Formulation of RS2GS and GS2RS

In this subsection, we propose a unified formulation of RS2GS and GS2RS so that Eqs. (1), (2), (3), *i.e.*, RS2GS and GS2RS conversions, can be simultaneously implemented using the same set of shared learnable parameters θ. In the following, based on the approximated bidirectional motion fields $\mathbf{F}_{0\to t}, \mathbf{F}_{1\to t}$, which are obtained by performing a simple linear scaling operation on the bidirectional optical flow fields $\mathbf{F}_{0\to 1}, \mathbf{F}_{1\to 0}$, the warping models of RS2GS and GS2RS can be established effectively.

RS2GS formulation. We note that the number of scanlines of the input image is H, and the pixel point $\mathbf{q}_i$ is located at y_i-th scanline, where $1 \le y_i \le H$. As shown in Fig. 3 (a), suppose that $\mathbf{q}_0$ in $\mathbf{I}_0^{\mathrm{RS}}$ and $\mathbf{q}_1$ in $\mathbf{I}_1^{\mathrm{RS}}$ are a pair

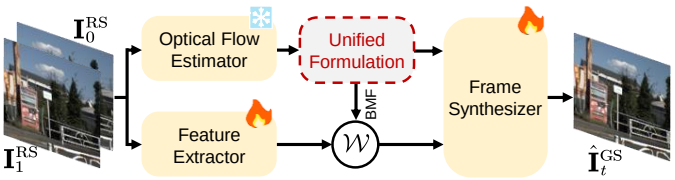

Fig. 5. Architecture overview of the proposed SelfRSSplat. Taking the RS2GS conversion in Eq. (1) as an example, given two consecutive RS images, our method consists of three main steps to recover a latent GS image $\hat{\mathbf{I}}_t^{\mathrm{GS}}$ corresponding to time t. First, contextual features are extracted; then, based on the estimated bidirectional optical flow, the bidirectional motion fields (BMF) are obtained using our unified formulation of RS2GS and GS2RS (*i.e.*, Eq. (6)); finally, the warped features are decoded into the target GS image in the frame synthesizer. In particular, this pipeline also accommodates Eq. (2) and Eq. (3), ensuring the feasibility of self-supervised cycle-consistency learning.

of matching points, which passes through an intermediate pixel point $\mathbf{q}_t$ in $\mathbf{I}_t^{\mathrm{GS}}$. Therefore, $\mathbf{F}_{0\to1}(\mathbf{q}_0) = \overrightarrow{\mathbf{I}_0^{\mathrm{RS}}(\mathbf{q}_0)\mathbf{I}_1^{\mathrm{RS}}(\mathbf{q}_1)}$ and $\mathbf{F}_{1\to0}(\mathbf{q}_1) = \overrightarrow{\mathbf{I}_1^{\mathrm{RS}}(\mathbf{q}_1)\mathbf{I}_0^{\mathrm{RS}}(\mathbf{q}_0)}$ can represent a set of bidirectional optical flow vectors. To obtain the bidirectional motion vectors $\mathbf{F}_{0\to t}(\mathbf{q}_0) = \overrightarrow{\mathbf{I}_0^{\mathrm{RS}}(\mathbf{q}_0)\mathbf{I}_t^{\mathrm{GS}}(\mathbf{q}_t)}$ and $\mathbf{F}_{1\to t}(\mathbf{q}_1) = \overrightarrow{\mathbf{I}_1^{\mathrm{RS}}(\mathbf{q}_1)\mathbf{I}_t^{\mathrm{GS}}(\mathbf{q}_t)}$, we try to employ the constraint of similar triangles. Specifically, we first approximate the scaling factor of RS2GS as $\mathbf{\Phi}_t^{\mathrm{GS}}(\mathbf{q}_0,\mathbf{q}_1) = \mathbf{F}_{0\to t}(\mathbf{q}_0)/\mathbf{F}_{0\to1}(\mathbf{q}_0)$, which can be expanded as:

$$\mathbf{\Phi}_t^{\mathrm{GS}}(\mathbf{q}_0,\mathbf{q}_1) = \frac{t+\gamma-\frac{\gamma}{H}y_0}{1+\frac{\gamma}{H}(y_1-y_0)}. \tag{5}$$

Note that, the vertical component of the forward optical flow vector $\mathbf{F}_{0\to1}(\mathbf{q}_0)$ can yield $y_1-y_0 = \mathbf{F}_{0\to1}^v(\mathbf{q}_0)$; the use of the backward optical flow vector $\mathbf{F}_{1\to0}(\mathbf{q}_1)$ results in $y_1-y_0 = -\mathbf{F}_{1\to0}^v(\mathbf{q}_1)$. Thus, the bidirectional scaling factors can be formulated as $\mathbf{\Phi}_t^{\mathrm{f}}(\mathbf{q}_0) \triangleq \mathbf{\Phi}_t^{\mathrm{GS}}(\mathbf{q}_0,\mathbf{q}_1)$, $\mathbf{\Phi}_t^{\mathrm{b}}(\mathbf{q}_1) \triangleq \mathbf{\Phi}_t^{\mathrm{GS}}(\mathbf{q}_1,\mathbf{q}_0)$. Then, stacking all pixels in matrix form, we can obtain the bidirectional scaling fields $\mathbf{\Phi}_t^{\mathrm{f}}$, $\mathbf{\Phi}_t^{\mathrm{b}}$. Finally, linearly scaling the bidirectional optical flow fields can generate the bidirectional motion fields as follows:

$$\begin{aligned}\mathbf{F}_{0\to t} &= \mathbf{\Phi}_t^{\mathrm{f}} \odot \mathbf{F}_{0\to1},\\ \mathbf{F}_{1\to t} &= (\mathbf{1}-\mathbf{\Phi}_t^{\mathrm{b}}) \odot \mathbf{F}_{1\to0},\end{aligned} \tag{6}$$

where $\odot$ is the Hadamard product. The theoretical proof of Eq. (6) can be found in [8], [17], involving the parallax effect caused by depth variation. Heretofore, we have completed the formulation of RS2GS, which can be utilized to warp the RS input to a virtual GS counterpart corresponding to time $t \in [-\gamma/2, 1-\gamma/2]$.

GS2RS formulation. It is worth noting that we cleverly propose to *construct the opposite time axis* (*vs.* RS imaging) to subtly model GS2RS, as illustrated in Fig. 3 (b). Analogously, $\mathbf{F}_{0\to t}(\mathbf{q}_0) = \overrightarrow{\mathbf{I}_0^{\mathrm{GS}}(\mathbf{q}_0)\mathbf{I}_t^{\mathrm{RS}}(\mathbf{q}_t)}$ and $\mathbf{F}_{1\to t}(\mathbf{q}_1) = \overrightarrow{\mathbf{I}_1^{\mathrm{GS}}(\mathbf{q}_1)\mathbf{I}_t^{\mathrm{RS}}(\mathbf{q}_t)}$ can denote a set of bidirectional motion vectors. To interpolate the RS frame $\mathbf{I}_t^{\mathrm{RS}}$ corresponding to time t from the two adjacent GS frames $\mathbf{I}_0^{\mathrm{GS}}$ and $\mathbf{I}_1^{\mathrm{GS}}$, we still exploit the linear motion assumption, often used in VFI [20], [21], [28], [34] and RSTSR [8], [9]. To begin with, we can easily obtain $||\overrightarrow{\mathbf{I}_0^{\mathrm{GS}}(\mathbf{q}_0)\mathbf{I}_t^{\mathrm{RS}}(\mathbf{q}_0)}|| = t+\gamma-\frac{\gamma}{H}y_0$, $||\overrightarrow{\mathbf{I}_1^{\mathrm{GS}}(\mathbf{q}_1)\mathbf{I}_t^{\mathrm{RS}}(\mathbf{q}_1)}|| = 1-t-\gamma+\frac{\gamma}{H}y_1$. Furthermore, using the similarity constraint on triangles, namely, $||\overrightarrow{\mathbf{I}_1^{\mathrm{GS}}(\mathbf{q}_1)\mathbf{I}_t^{\mathrm{RS}}(\mathbf{q}_1)}|| \cdot ||\mathbf{F}_{0\to t}(\mathbf{q}_0)|| = ||\overrightarrow{\mathbf{I}_0^{\mathrm{GS}}(\mathbf{q}_0)\mathbf{I}_t^{\mathrm{RS}}(\mathbf{q}_0)}|| \cdot ||\mathbf{F}_{1\to t}(\mathbf{q}_1)||$, the scaling factor of GS2RS can be approximated as $\mathbf{\Phi}_t^{\mathrm{RS}}(\mathbf{q}_0,\mathbf{q}_1) = \mathbf{F}_{0\to t}(\mathbf{q}_0)/\mathbf{F}_{0\to1}(\mathbf{q}_0)$, *i.e.*,

$$\mathbf{\Phi}_t^{\mathrm{RS}}(\mathbf{q}_0,\mathbf{q}_1) = \frac{t+\gamma-\frac{\gamma}{H}y_0}{1+\frac{\gamma}{H}(y_1-y_0)}. \tag{7}$$

It can be seen that Eq. (7) and Eq. (5) share the same structure despite serving different conversion purposes. We thus can similarly derive the bidirectional scaling fields $\mathbf{\Phi}_t^{\mathrm{f}}$, $\mathbf{\Phi}_t^{\mathrm{b}}$ based on Eq. (7), as in Eq. (5). Subsequently, the bidirectional motion fields $\mathbf{F}_{0\to t}$ and $\mathbf{F}_{1\to t}$ can be obtained by linearly scaling the regular bidirectional optical flow fields. Note that this GS2RS process can also be formulated by Eq. (6), consistent with RS2GS. Hence, referring to Fig. 2, inputting $\hat{\mathbf{I}}_{t+1}^{\mathrm{GS}}$ first and then $\hat{\mathbf{I}}_t^{\mathrm{GS}}$ in Eq. (3) can be used to reconstruct back the original RS frame $\hat{\mathbf{I}}_1^{\mathrm{RS}}$, which is the key to achieve self-supervised cycle consistency learning. At this time, the formulation of GS2RS naturally remains the same as the above RS2GS formulation, that is, the unification of GS2RS and RS2GS warping models is achieved.

D. Architecture of the Proposed SelfRSSplat

Building upon the aforementioned unified formulation in Eq. (6), we would like to have a uniform network model at hand, which can be compatible with both RS2GS and GS2RS conversions like $\mathcal{M}$ in Eqs. (1), (2), (3). To this end, as shown in Fig. 5, we directly inject the parameter-free unified model of RS2GS and GS2RS into the motion estimation module of the well-established VFI method SoftSplat [21], which can seamlessly adapt to both RS2GS and GS2RS conversions, named **RSSplat**. The feature extractor and frame synthesizer of RSSplat follow the structure of SoftSplat. Specifically, it first estimates the bidirectional optical flow fields by the pre-trained GMFlow [71], then performs forward warping of the RS (or GS) context features in accordance with the bidirectional motion fields approximated by Eq. (6), and finally combines and refines the warped features into the target GS (or RS) image within a frame synthesizer. In this way, motion compensation and occlusion reasoning can be carried out efficiently.

Note that RSSplat inherits the advantages of using feature-based synthesis like prevailing RS correction methods [7], [10], [11], [18], [19] (*cf.*, Table I) and VFI [21], [34], [50], [51], [53] methods, and exhibits superior performance compared to the image-based synthesis method CVR [9]. We also believe extending the well-established VFI method (*i.e.*, SoftSplat) to support RS2GS and GS2RS can provide a valuable perspective. Ultimately, based on our self-supervised learning framework in Sec. III-B, the model parameters θ of RSSplat can be optimized without using GT RS-GS image pairs, yielding our self-supervised RSTSR model, **SelfRSSplat**, to unroll two consecutive RS images into a smooth GS sequence.

E. Loss Function

The proposed pipeline can be end-to-end trained in a self-supervised learning manner by minimizing the discrepancy

TABLE II
QUANTITATIVE RESULTS AT TIME $1-\gamma$ ON CARLA-RS AND FASTEC-RS DATASETS. GRAY PART INDICATES THE TIME-SPECIFIC RS CORRECTION METHOD AND THE REMAINDER IS THE TIME-ARBITRARY RSTSR METHOD. **BOLD** AND UNDERLINED NUMBERS DENOTE THE BEST AND SECOND-BEST PERFORMANCE IN RSTSR METHODS. OUR SELF-SUPERVISED PIPELINE ACHIEVES COMPARABLE OR BETTER PERFORMANCE THAN SUPERVISED METHODS.

Method	Supervision	#Params (Million)	PSNR↑ (dB)			SSIM↑		LPIPS↓	
			CRM	CR	FR	CR	FR	CR	FR
DiffSfM [69]	✗	-	24.20	21.28	20.14	0.775	0.701	0.1322	0.1789
CIEUnroll [16]	✓	-	31.84	31.43	28.57	0.919	0.844	-	-
DSUN [10]	✓	3.91	26.90	26.46	26.52	0.807	0.792	0.0703	0.1222
SUNet [11]	✓	12.0	29.28	29.18	28.34	0.850	0.837	0.0658	0.1205
DSUN [10] + BMBC [34]	✓	14.9	27.29	27.58	24.95	0.829	0.787	0.0980	0.2024
DSUN [10] + DAIN [51]	✓	27.9	27.48	27.88	26.19	0.874	0.807	0.0821	0.1453
JAMNet [19] + SoftSplat [21]	✓	12.1	30.40	30.14	26.63	0.895	0.815	0.0629	0.1982
JAMNet [19] + RIFE [22]	✓	15.4	29.96	29.74	26.81	0.877	0.813	0.1241	0.2315
RSSR [8]	✓	26.0	30.17	24.78	21.23	0.867	0.776	0.0695	0.1659
CVR [9]	✓	42.7	32.02	31.74	28.72	0.929	0.847	0.0368	0.1107
RSSplat (Ours)	✓	**7.44**	**32.92**	**32.80**	**29.85**	**0.938**	**0.869**	**0.0197**	**0.0814**
SelfRSSplat (Ours)	✗	**7.44**	31.06	31.00	27.25	0.915	0.814	0.0260	0.0828

between the input middle RS frame $\mathbf{I}_1^{\text{RS}}$ and its cycle reconstruction $\hat{\mathbf{I}}_1^{\text{RS}}$. In addition, possible color distortion of intermediate GS predictions $\hat{\mathbf{I}}_t^{\text{GS}}, \hat{\mathbf{I}}_{t+1}^{\text{GS}}$ needs to be handled during cycle consistency training. As a consequence, our total loss function incorporates cycle consistency loss $\mathcal{L}_{\text{cycle}}$ and color consistency loss $\mathcal{L}_{\text{color}}$, *i.e.*, $\mathcal{L} = \mathcal{L}_{\text{cycle}} + \mathcal{L}_{\text{color}}$.

Cycle consistency loss $\mathcal{L}_{\text{cycle}}$. It is a linear combination of image reconstruction loss and perceptual loss [72]. The former measures the pixel-wise cycle reconstruction error, and the latter contributes to preserving fine details and enhancing the perceptual quality, given by

$$\mathcal{L}_{\text{cycle}} = \rho(\mathbf{I}_1^{\text{RS}} - \hat{\mathbf{I}}_1^{\text{RS}}) + \lambda \|\phi(\mathbf{I}_1^{\text{RS}}) - \phi(\hat{\mathbf{I}}_1^{\text{RS}})\|_1, \tag{8}$$

where $\rho(x) = \sqrt{x^2 + \varepsilon^2}$ is the Charbonnier function with constant ε of 0.001, ϕ is the conv4_3 feature of the VGG16 model [73], and the weight λ is empirically set to 0.01.

Color consistency loss $\mathcal{L}_{\text{color}}$. As illustrated in Fig. 4, the estimated intermediate GS frames $\hat{\mathbf{I}}_t^{\text{GS}}, \hat{\mathbf{I}}_{t+1}^{\text{GS}}$ from Eqs. (1), (2) may suffer from color distortion. Although this does not affect the cycle reconstruction of $\hat{\mathbf{I}}_1^{\text{RS}}$ in Eq. (3), it will lead to degradation of the resulting SelfRSSplat model during inference (*i.e.*, Eq. (1)). Therefore, we force the warped intermediate GS frame to align with the input middle RS frame $\mathbf{I}_1^{\text{RS}}$ based on temporal distance, *i.e.*,

$$\mathcal{L}_{color} = \begin{cases} \rho(\mathbf{I}_1^{\text{RS}} - \mathcal{W}(\hat{\mathbf{I}}_{t+1}^{\text{GS}}, \mathbf{F}_{0\to t}^{\text{GS2RS}})) & \text{if } t \le (1-\gamma)/2 \\ \rho(\mathbf{I}_1^{\text{RS}} - \mathcal{W}(\hat{\mathbf{I}}_t^{\text{GS}}, \mathbf{F}_{1\to t}^{\text{GS2RS}})) & \text{if } t > (1-\gamma)/2 \end{cases} \tag{9}$$

where $\mathcal{W}$ indicates the frame warping operator, and $\mathbf{F}_{0\to t}^{\text{GS2RS}}$, $\mathbf{F}_{1\to t}^{\text{GS2RS}}$ refer to the bidirectional motion fields approximated during the execution of Eq. (3). We found that such a simple constraint is sufficient for SelfRSSplat to maintain the correct color in the RGB space.

IV. EXPERIMENTS

A. Experimental Setup

Datasets. We adopt three standard RS benchmark datasets, including Carla-RS [10], Fastec-RS [10], and BS-RSC [7]. The synthetic Carla-RS dataset is generated from a virtual 3D environment, involving general 6-DoF camera motions, and the image resolution is 640×448. The Fastec-RS dataset consists of RS image sequences synthesized by row-by-row

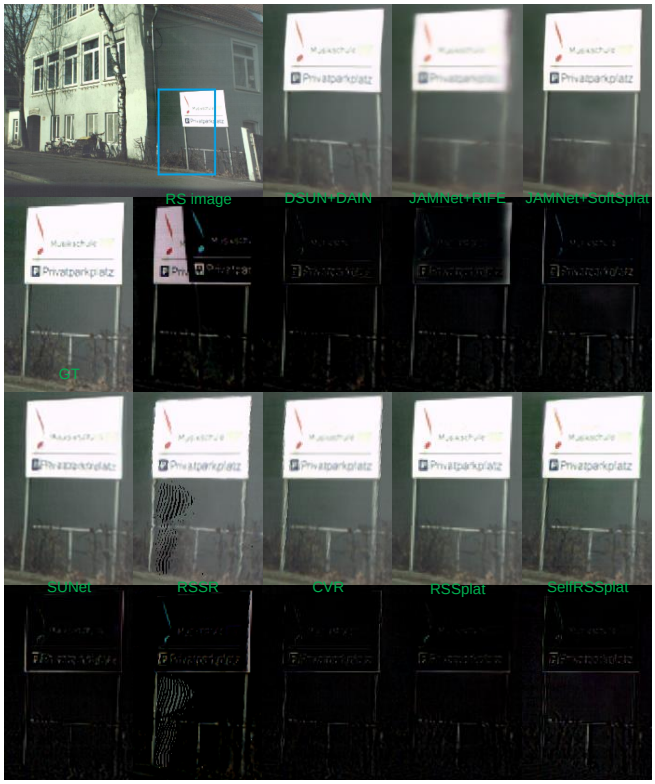

Fig. 6. Visual comparison of RS correction results on the Fastec-RS dataset [10]. Even rows: the absolute difference between the recovered image and the corresponding GT GS image.

stitching of high-speed GS videos with dynamic scenes, with an image resolution of 640×480. Note that they provide GS GT corresponding to the first and middle scanlines, and the readout time ratios are all 1. Thus, GT RS-GS image pairs with $t = \{-0.5, 0, 0.5\}$ can be used for supervised training. The real-world BS-RSC dataset [7] released later contains a variety of camera and object movements (*e.g.*, vehicles and pedestrians) in urban environments, with a readout time ratio γ of 0.45 and an image resolution of 1024×768. Note that only the middle-scanline GS GT is collected by a beam-splitter system. Due to limited GS GT being recorded, it could previously only be suitable for time-specific RS correction [7], [17], [19] and not for time-arbitrary RSTSR tasks. In contrast, our SelfRSSplat

TABLE III
QUANTITATIVE RESULTS AT TIME $1-\gamma/2$ ON CARLA-RS AND FASTEC-RS DATASETS. GRAY PART INDICATES THE TIME-SPECIFIC RS CORRECTION METHOD AND THE REMAINDER IS THE TIME-ARBITRARY RSTSR METHOD. **BOLD** AND UNDERLINED NUMBERS DENOTE THE BEST AND SECOND-BEST PERFORMANCE IN TIME-ARBITRARY RSTSR METHODS. IN ADDITION TO THE OUTSTANDING PERFORMANCE AT TIME $1-\gamma$ IN TABLE II, THE SUPERIOR PERFORMANCE AT TIME $1-\gamma/2$ ALSO PROVES THE ADVANTAGE OF OUR PIPELINE IN HIGH-FRAME-RATE GS VIDEO RECONSTRUCTION.

Method	Supervision	PSNR↑ (dB)			SSIM↑		LPIPS↓	
		CRM	CR	FR	CR	FR	CR	FR
DiffSfM [69]	✗	25.93	22.88	21.44	0.770	0.710	0.1201	0.2180
AdaRSC [7]	✓	-	-	28.56	-	0.855	-	0.0796
JCD [18]	✓	28.12	27.75	26.48	0.836	0.821	0.0595	0.0943
DSUN [10]	✓	27.86	27.54	26.73	0.829	0.819	0.0555	0.0995
SUNet [11]	✓	28.44	28.17	27.06	0.838	0.825	0.0702	0.1030
JAMNet [19]	✓	31.00	30.70	28.70	0.905	0.865	0.0371	0.0691
SUNet [11] + BMBC [34]	✓	28.51	28.69	25.49	0.848	0.796	0.1033	0.2118
SUNet [11] + DAIN [51]	✓	28.63	28.93	27.12	0.851	0.823	0.0919	0.1642
RSSR [8]	✓	29.36	26.57	24.89	0.900	0.824	0.0553	0.1109
CVR [9]	✓	29.41	29.19	26.67	0.915	0.838	0.0403	0.1011
RSSplat (Ours)	✓	**31.93**	**31.75**	**28.98**	**0.929**	**0.864**	**0.0222**	0.0797
SelfRSSplat (Ours)	✗	29.49	29.35	26.47	0.894	0.810	0.0292	**0.0684**

enables the recovery of time-arbitrary GS images in this dataset for the first time.

Training details. Since the optical flow estimation model pre-trained on GS images can be used on RS images [2], [17], [24], [69], we employ a pre-trained optical flow estimator and keep it frozen, training only the feature extractor and frame synthesizer, as shown in Fig. 5. Note that SelfRSSplat is trained within our self-supervised learning framework, whereas RSSplat is used as an RSTSR model with supervised training from scratch. The model is trained for 200 epochs using the Adam optimizer with a learning rate of 0.0001 and a batch size of 2. We randomly sample diverse $t \in [-\gamma/2, 1-\gamma/2]$ in the Carla-RS and Fastec-RS datasets, and fix t to $1-\gamma$ in the BS-RSC dataset, since 55% blank rows between adjacent RS frames caused by γ of 0.45 may lead to training instability (*cf.*, Sec. IV-F). 320×256 patches are cropped randomly from a triplet of RS images. Random horizontal flipping and vertical flipping combined with reverse order [19], are introduced for data augmentation. All experiments are performed on a single NVIDIA RTX 3090 GPU. At the test phase, our method can reconstruct GS frames corresponding to any time $t \in [-\gamma/2, 1-\gamma/2]$.

Evaluation protocols. Since GT occlusion mask is available in the Carla-RS dataset, following [9], [10], [24], we perform the evaluation as: Carla-RS with mask (*CRM*), Carla-RS without mask (*CR*), and Fastec-RS (*FR*). Standard PSNR and SSIM metrics, and learned perceptual metric LPIPS [74] are applied. A higher PSNR/SSIM (↑) or lower LPIPS (↓) score indicates better performance.

Comparison methods. We evaluate our method against the following baselines: (1) **DiffSfM** [69], **CIEUnroll** [16], **JCD** [18], **AdaRSC** [7], **JAMNet** [19], **DSUN** [10], **SUNet** [11] are time-specific RS correction methods. (2) **RSSR** [8] and **CVR** [9] are popular RSTSR methods that can restore time-arbitrary GS images from two adjacent RS images. Since Qu *et al.* [24] requires five consecutive RS frames as input, for fairness, we do not compare with it. (3) **Cascaded methods** contain four paradigms of RS correction followed by VFI methods, *i.e.*, "DSUN + BMBC [34]", "DSUN + DAIN [51]", "JAMNet + SoftSplat [21]", and "JAMNet + RIFE [22]".

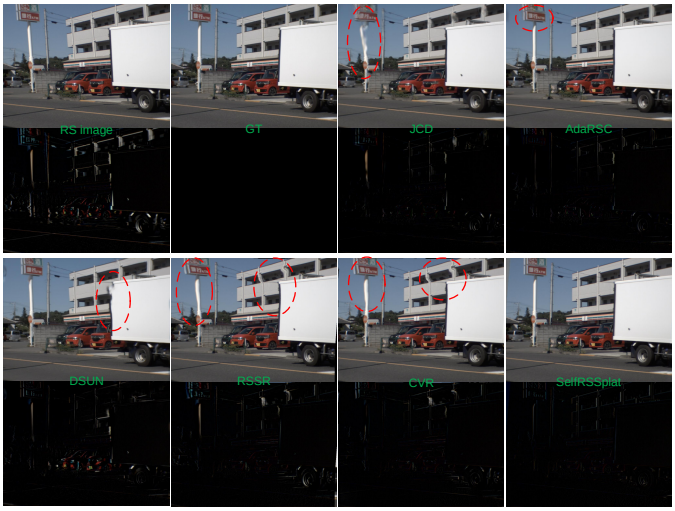

Fig. 7. Visual comparison of RS correction results on the real-world BS-RSC dataset [7]. Areas with significant distortion are marked with red circles. Even rows: the absolute difference between the recovered image and the corresponding GT GS image.

B. Comparison and Analysis

1) Results on Carla-RS and Fastec-RS datasets: The quantitative results corresponding to time $1-\gamma$ are reported in Table II. Thanks to the context synthesis in feature space, RSSplat consistently achieves SoTA results. Without using GS GT references for supervision, our SelfRSSplat shows competitive performance with prior supervised methods and even exceeds them on the LPIPS metric, demonstrating the effectiveness of our proposed self-supervised learning framework. Additionally, our approach features a more lightweight network architecture.

Table III shows the RS correction result corresponding to time $1-\gamma/2$. It can also be seen that RSSplat (supervised) benefits from the powerful feature-based synthesis capability and achieves significantly better results than the image-based synthesis method CVR (supervised). Note that due to SelfRSSplat's ability to leverage the intrinsic distribution information of the input RS images themselves, it holds the potential to surpass current supervised baselines in terms of visual perceptual metrics, making the GS recovery results more faithful to the latent GS image.

Fig. 8. Visual comparison with existing RSTSR methods (*i.e.*, RSSR and CVR) on real-world BS-RSC dataset [7]. Six consecutive latent GS video frames are recovered from two adjacent RS frames. Moving pedestrians and vehicles are included. Fluid GS video sequences cannot be generated by supervised methods RSSR and CVR because the limited GS GT is not efficiently sufficient to complete frame synthesis in RGB color space. Our SelfRSSplat circumvents this problem by means of self-supervised learning to recover temporally consistent and visually pleasant GS video sequences.

A visual example is displayed in Fig. 6, where our method exhibits impressive results in terms of local details and subjective perception even though it is self-supervised. Due to error accumulation, cascaded methods are prone to blurring artifacts and local inaccuracies. For example, JAMNet+SoftSplat and JAMNet+RIFE have obvious unclear artifacts in grasses and railings behind. These experiments collectively underscore the effectiveness and superiority of our proposed self-supervised architecture in RS effect removal.

2) Results on BS-RSC dataset: As shown in Table IV and Fig. 7, our method eliminates the dependence on GS GT reference and performs effectively in real-world scenarios, even surpassing most time-specific RS correction methods. Moreover, our SelfRSSplat achieves a *2.39 dB* PSNR improvement compared to SoTA time-arbitrary method CVR [9]. Note that our pipeline also opens new avenues to unroll two consecutive RS images into a high frame rate GS video for this dataset, as evident in Sec. IV-C.

TABLE IV
QUANTITATIVE COMPARISON ON THE REAL-WORLD BS-RSC DATASET [7]. DUE TO THE LIMITED GS GT PROVIDED, IT IS NOT ENOUGH TO TRAIN RELIABLE RSTSR MODELS IN A SUPERVISED MANNER. OUR APPROACH IS THE FIRST TO RECOVER TIME-ARBITRARY GS FRAMES FOR THIS DATASET.

Method	PSNR↑ (dB)	SSIM↑	Time-Arbitrary
DSUN [10]	25.21	0.833	✗
SUNet [11]	27.76	0.875	✗
JCD [18]	25.59	0.841	✗
AdaRSC [7]	28.23	0.882	✗
JAMNet [19]	**32.93**	**0.941**	✗
RSSR [8]	26.47	0.880	✗
CVR [9]	28.14	0.895	✗
SelfRSSplat (Ours)	30.53	0.914	✓

C. RS Temporal Super-Resolution Results

It should be emphasized that our approach not only outperforms time-specific RS correction methods but also adapts to time-arbitrary RSTSR tasks, *i.e.*, our method can recover intermediate GS frames corresponding to arbitrary time steps. The visual result of $5\times$ temporal upsampling on the real-world BS-RSC dataset is illustrated in Fig. 8. Furthermore, we present a *supplementary video* to demonstrate the RSTSR results.

Due to hardware limitations, the BS-RSC dataset only collects limited GS GT, which is insufficient for supervised training of RSTSR networks. As a result, previous supervised methods RSSR and CVR face difficulties in recovering smooth temporal sequences of latent GS images from two consecutive RS frames. For example, the pillar in Fig. 8 cannot be correctly corrected in GS images far from the GT supervisory signal. In addition, our approach is robust to moving objects, *e.g.*, pedestrians and vehicles in Fig. 8. We conjecture this is because the GS frame candidates obtained by the linear motion-based warping can be effectively fused in the frame synthesizer. Benefiting from

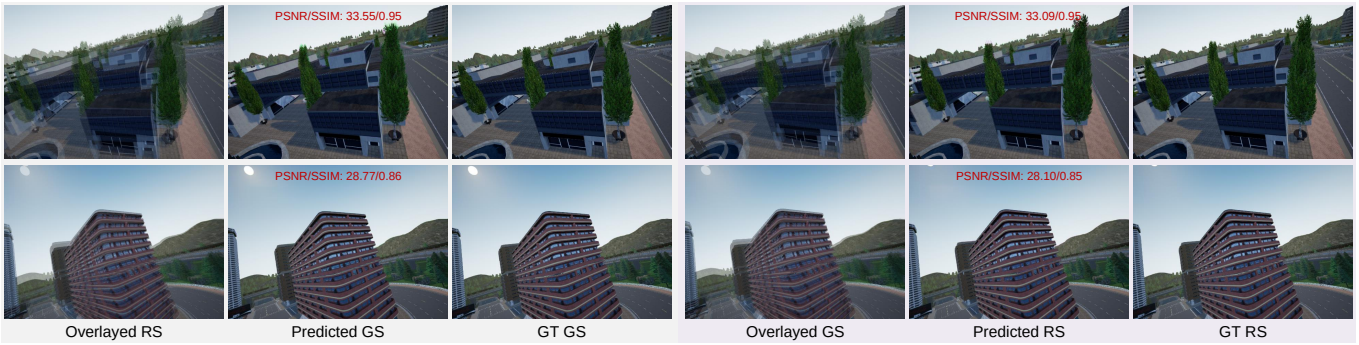

Fig. 9. Qualitative results of RS2GS and GS2RS conversions on the Carla-RS dataset. Our method not only supports RS correction to restore the underlying GS image (*i.e.*, RS2GS on the left), but can also be used in turn to synthesize a high-quality RS image (*i.e.*, GS2RS on the right). Note that they share the same network model of SelfRSSplat.

TABLE V
QUANTITATIVE RESULTS OF GS2RS AND RS2GS CONVERSIONS. OUR APPROACH ALLOWS EFFICIENT INTERCONVERSION BETWEEN RS IMAGES AND GS IMAGES USING THE SAME NETWORK MODEL.

Conversion	PSNR↑ (dB)			SSIM↑		LPIPS↓	
	CRM	CR	FR	CR	FR	CR	FR
GS2RS	30.98	30.95	27.14	0.914	0.809	0.0268	0.0842
RS2GS	31.06	31.00	27.25	0.915	0.814	0.0260	0.0828

the proposed self-supervised learning framework, our approach not only successfully eliminates RS artifacts and preserves rich image details, but also effectively reconstructs temporally continuous GS sequences for various real-world RS images.

D. GS2RS Conversion Results

An important advantage of our proposed pipeline is its self-supervised nature, which benefits from the widely-used cycle consistency constraint and our proposed unified formulation of RS2GS and GS2RS. Another advantage is its common compatibility with both RS2GS and GS2RS conversions. This nature is due to the unification of RS2GS and GS2RS warping models in our formulation in Sec. III-C, allowing a common network model to adapt to the two complement conversion attributes at the same time.

As shown in Table V and Fig. 9, our method can be applied not only to RS2GS conversion, *i.e.*, generating high-frame-rate and high-fidelity GS videos from two adjacent RS frames, but also to GS2RS conversion. Note that GS2RS conversion can potentially provide a promising possibility for simulating RS datasets from real low-frame-rate GS videos. Previously, a naive approach to this goal at hand would be synthesizing high-frame-rate GS videos from low-frame-rate GS videos by off-the-shelf video frame interpolation (VFI) methods [20], [22], [51], and then performing row-by-row stitching, such as [16], [25], [26]. Nevertheless, this two-stage process is time-consuming and labor-intensive, whereas our GS2RS conversion is capable of efficiently doing it in one step. It is also worth emphasizing that the GS2RS conversion is a key element for constructing the cycle consistency constraint, which has not been investigated before. Driven by the mutual conversion between RS2GS and GS2RS, a tractable self-supervised learning framework is built effectively in this paper.

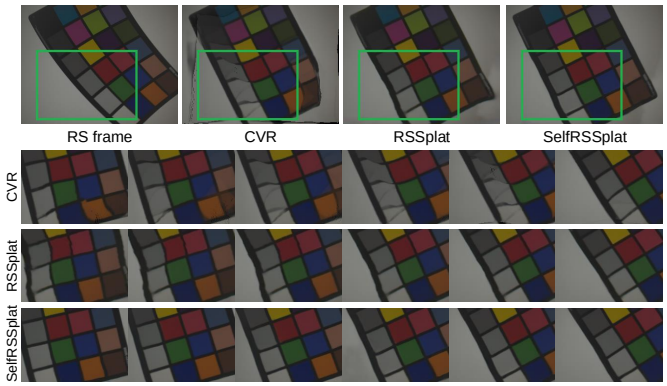

Fig. 10. Generalization comparison of RSTSR on Gev-RS-Real test data [26]. For severe RS distortions caused by a fast-swinging calibration plate in real-world scenarios, supervised CVR and RSSplat struggle to generalize to interpolate a reliable GS sequence, while our SelfRSSplat delivers smooth GS videos by virtue of self-supervised learning.

E. Generalization Evaluation on Other Real Data

To evaluate the generalization of our method, we adopt the Gev-RS-Real test data [26], which contains noticeable real-world RS distortions and the GS GT is not available. The image resolution is 346×260. We utilize the model pre-trained on the real-world BS-RSC dataset due to smaller domain gaps. As shown in Fig. 10, supervised CVR and RSSplat methods fall short when dealing with new real RS images due to significant differences in intermediate motion estimation, *i.e.*, a large number of visual artifacts and temporal discontinuities are generated. In contrast, our SelfRSSplat directly fits the real-world data distribution to model intermediate motion relations in a self-supervised fashion, resulting in visually fluid and temporally coherent results and reinforcing the potential of RS cameras for broader applications.

Moreover, we also adopt the real RS data provided by [69] and [75], in which a mobile phone is held and moved quickly and irregularly outdoors, resulting in significant RS distortions. The image resolutions are 1280×720 and 640×480 respectively. The qualitative RS correction results are displayed in Fig. 11. It can be seen that our method is able to effectively and robustly remove real-world RS effects. While there is no access to GS GT references *e.g.*, for most real RS devices, our SelfRSSplat can more comprehensively explore the intermediate motion modeling of real RS data, thereby recovering temporally

Fig. 11. RS correction results on real-world RS images provided by [69] (left three columns) and [75] (right three columns). Our method can robustly generalize to remove significant RS artifacts in practical scenarios.

TABLE VI
ABLATION STUDIES ON SUPERVISION TIME, SYNTHESIS STRATEGY, AND TRAINING APPROACH. "DIVERSE t" MEANS THE GS RECONSTRUCTION TIME t IS RANDOMLY SAMPLED IN THE TIME INTERVAL $[-\gamma/2, 1-\gamma/2]$ DURING TRAINING. FIXED t MEANS THAT THE GS RECONSTRUCTION TIME t IS FIXED TO $1-\gamma$ DURING TRAINING. SELFCVR DENOTES THE CVR NETWORK [9] RETRAINED WITHIN OUR SELF-SUPERVISED LEARNING FRAMEWORK. "W/O FROZEN" INDICATES THAT THE PARAMETERS OF THE OPTICAL FLOW ESTIMATOR ARE NOT FROZEN DURING SELF-SUPERVISED TRAINING. † INDICATES TRAINING INSTABILITY. SINCE THE BS-RSC DATASET HAS 55% INTER-FRAME BLANK ROWS, THE SELF-SUPERVISED TRAINING WITH IMAGE-BASED SYNTHESIS OR DIVERSE t MAY SUFFER FROM INSTABILITY OR FAILURE.

Method	Carla-RS dataset			Fastec-RS dataset			BS-RSC dataset	
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM
SelfRSSplat (fixed t)	29.35	0.881	0.0261	25.93	0.789	0.0916	**30.53**	**0.914**
SelfRSSplat (diverse t)	**31.00**	**0.915**	**0.0260**	**27.25**	**0.814**	**0.0828**	28.41†	0.902†
SelfCVR (image-based)	30.78	**0.915**	0.0398	26.70	0.804	0.1111	24.86†	0.820†
SelfRSSplat (feature-based)	**31.00**	**0.915**	**0.0260**	**27.25**	**0.814**	**0.0828**	**30.53**	**0.914**
SelfRSSplat (w/o frozen)	30.08	0.888	0.0375	26.21	0.787	0.1055	29.84	0.901
SelfRSSplat (w/ frozen)	**31.00**	**0.915**	**0.0260**	**27.25**	**0.814**	**0.0828**	**30.53**	**0.914**

consistent and geometrically correct distortion-free images. These experiments verify that our method enjoys strong generalization performance in practice, which also expands more possibilities for the actual use of RS cameras.

F. Ablation Studies

We conduct ablation studies concerning supervision time, synthesis strategy, training approach, and network structure.

Ablations on supervision time. We separately utilize diverse $t \in [-\gamma/2, 1-\gamma/2]$ and fixed $t = 1-\gamma$ for network training. Since Carla-RS and Fastec-RS datasets have good temporal continuity ($\gamma = 1$), as shown in Table VI, sampling diverse t during training is conducive to superior results. By contrast, the BS-RSC dataset with $\gamma = 0.45$ creates 55% blank rows across adjacent RS frames. This longer temporal distance makes training based on diverse t prone to instability. Hence, we make t fixed during the self-supervised training of this dataset, which not only facilitates more stable network training but also yields more accurate recovery results. And thanks to feature-based synthesis as in [21], this does not prevent our method from producing a continuous GS video.

TABLE VII
EFFICACY UNDER DIFFERENT OPTICAL FLOW ESTIMATORS (GMFLOW [71] AND RAFT [43]). PSNR/SSIM AND NETWORK PARAMETERS FOR SELFRSSPLAT ARE REPORTED. RUNTIME IS TESTED ON 640×480 IMAGES AND A SINGLE 3090 GPU.

SelfRSSplat		w/ RAFT [43]	w/ GMFlow [71]
Dataset	Carla-RS	**31.72/0.929**	31.00/0.915
	Fastec-RS	**28.13/0.832**	27.25/0.814
	BS-RSC	**30.85/0.915**	30.53/0.914
Metric	Para. (Million)	8.02	**7.44**
	Time (ms)	201	**67**

Ablations on synthesis strategy. From the comparison of our RSSplat and CVR in Table II, feature-based synthesis achieves better results than image-based synthesis in the case of supervised training. To further understand its superiority in the context of self-supervised training, we implement a self-supervised version of CVR, called **SelfCVR**, based on our self-supervised learning framework. As can be seen from Table VI, our feature-based synthesis strategy exhibits more excellent RS correction performance. Note that SelfCVR is highly susceptible to training instability or even training failure, whether it is based on diverse t or fixed t, due to the presence of 55% blank rows between neighboring RS frames in the BS-

Fig. 12. Comparison of different synthesis strategies under supervised and self-supervised training. Compared with SelfCVR and CVR which perform frame synthesis in RGB color space, our feature-based synthesis methods SelfRSSplat and RSSplat exhibit better RS correction results.

Fig. 13. RS temporal super-resolution comparison of SelfCVR and SelfRSSplat on real-world GT-limited BS-RSC dataset. Six consecutive latent GS video frames are recovered from $\mathbf{I}_0^{RS}$ and $\mathbf{I}_1^{RS}$. Due to 55% inter-frame blank rows, the image-based synthesis method SelfCVR cannot produce a smooth and accurate GS video sequence, while our SelfRSSplat can synthesize a temporally and visually coherent GS video in the feature space.

RSC dataset. As shown in Fig. 12, our feature-based synthesis methods (*i.e.*, SelfRSSplat and RSSplat) have the potential to recover higher quality GS frames than image-based synthesis methods (*i.e.*, SelfCVR and CVR), whether in fully supervised or self-supervised contexts. This is also the reason why we base our self-supervised approach on RSSplat instead of CVR.

Furthermore, it is difficult for SelfCVR to recover a trustworthy and temporally continuous GS video, as illustrated in Fig. 13. Note that state-of-the-art supervised RSTSR methods RSSR and CVR also struggle to reconstruct a coherent GS video from two consecutive RS images in the BS-RSC dataset (*cf.*, Sec. IV-C). In contrast, our SelfRSSplat can effectively reconstruct continuous and smooth GS videos in a self-supervised learning manner even without GT GS images as the supervision reference, unlocking the potential of RS cameras for practical applications.

Ablations on optical flow estimator. Our approach employs a pre-trained optical flow estimator, with its parameters frozen during training. We first validate the effectiveness of parameter freezing. As evident from Table VI, freezing the optical flow estimator contributes to a more stable training process and superior GS reconstruction results. Subsequently, Table VII demonstrates the efficacy of our proposed self-supervised RSTSR framework when utilizing different optical flow estimators, namely GMFlow [71] and RAFT [43]. Our method effectively accommodates various optical flow estimators, attesting to its scalability. The adoption of GMFlow achieves a favorable balance between performance and efficiency. Consequently, advancements in future optical flow estimation methods are poised to benefit our framework as well.

G. Limitation and Discussion

As we all know, the more severe the relative motion between the RS camera and the scene, the more obvious the RS effect will be in the collected RS images. Consequently, in some typical applications, RS images with significant RS artifacts are often accompanied by motion blur. However, our SelfRSSplat employs GMFlow [71] to estimate the bidirectional optical flow fields between two clear RS images, and then utilizes the warped RS features for frame synthesis, both of which have limited adaptability and robustness to motion blur. Two examples on the BS-RSCD dataset [18] are shown

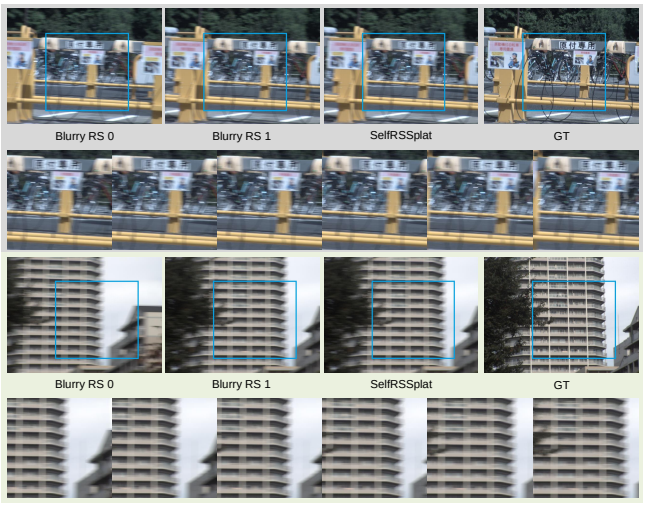

Fig. 14. Limitations of our method in processing blurry RS images in the BS-RSCD dataset [18]. For slightly blurry RS inputs, although our method can hallucinate a seemingly plausible continuous video sequence (see even rows), blurring artifacts are difficult to remove, which needs to be addressed in a targeted manner in the future.

in Fig. 14. For blurry RS input, although our method can generate a plausible latent video sequence, eliminating motion blur artifacts remains challenging. This is a common issue for current time-specific RS correction [7], [10], [11], [16], [19] and time-arbitrary RSTSR [8], [9], [24], [76] methods. We plan to handle this challenge in the future. Furthermore, how to extend our self-supervised learning framework to allow more consecutive RS frame inputs, which will bring better adaptability to non-linear motion like [7], [24], is also an interesting future research direction. Finally, constructing datasets with more ground truth GS images, along with corresponding camera motion parameters, would be valuable directions for comprehensively validating the quality of the reconstructed GS video and investigating the impact of different motion intensities.

V. Conclusion

In this paper, we have developed a self-supervised learning method that enables the recovery of intermediate GS frames with an arbitrary frame rate from two consecutive RS frames in the absence of GT RS-GS image pairs. Our approach opens up the possibility of learning self-supervised RS temporal super-resolution, while also incorporating advantages such as motion compensation and context aggregation. Importantly, spatio-temporal coherence is explored in our unified RS2GS and GS2RS formulation, ensuring the reuse of the same network model in RS2GS and GS2RS conversions. On this basis, a cycle consistency constraint is constructed from a triplet of consecutive RS frames, where two in-between GS frames are predicted and then leveraged to reconstruct the original middle RS frame. Experiments have been conducted to validate the effectiveness and excellent generalization capability of our approach on both synthetic and real data. It is hoped that this study can shed light for future research on self-supervised video reconstruction of RS cameras.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (Nos. 62136001, 62276007, 62088102, 62271410, 62401021). Bin Fan is sponsored by China National Postdoctoral Program for Innovative Talents (No. BX20230013) and China Postdoctoral Science Foundation (No. 2024M750101).

References

[1] J. Hedborg, P.-E. Forssén, M. Felsberg, and E. Ringaby, "Rolling shutter bundle adjustment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1434–1441.
[2] S. Vasu, M. M. Mohan, and A. Rajagopalan, "Occlusion-aware rolling shutter rectification of 3D scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 636–645.
[3] O. Saurer, K. Koser, J.-Y. Bouguet, and M. Pollefeys, "Rolling shutter stereo," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 465–472.
[4] Y. Lao, O. Ait-Aider, and A. Bartoli, "Solving rolling shutter 3D vision problems using analogies with non-rigidity," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 100–122, 2021.
[5] Y. Lu, G. Liang, and L. Wang, "Learning INR for event-guided rolling shutter frame correction, deblur, and interpolation," *arXiv preprint arXiv:2305.15078*, 2023.
[6] B. Fan, Y. Dai, and M. He, "Rolling shutter camera: Modeling, optimization and learning," *Machine Intelligence Research*, vol. 20, no. 6, pp. 783–798, 2023.
[7] M. Cao, Z. Zhong, J. Wang, Y. Zheng, and Y. Yang, "Learning adaptive warping for real-world rolling shutter correction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 785–17 793.
[8] B. Fan and Y. Dai, "Inverting a rolling shutter camera: Bring rolling shutter images to high framerate global shutter video," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 4228–4237.
[9] B. Fan, Y. Dai, Z. Zhang, Q. Liu, and M. He, "Context-aware video reconstruction for rolling shutter cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 572–17 582.
[10] P. Liu, Z. Cui, V. Larsson, and M. Pollefeys, "Deep shutter unrolling network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5941–5949.
[11] B. Fan, Y. Dai, and M. He, "SUNet: Symmetric undistortion network for rolling shutter correction," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 4541–4550.
[12] Y. Lao and O. 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 (CVPR)*, 2018, pp. 4795–4803.
[13] W. Yan, R. T. Tan, B. Zeng, and S. Liu, "Deep homography mixture for single image rolling shutter correction," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023, pp. 9868–9877.
[14] V. Rengarajan, Y. Balaji, and A. Rajagopalan, "Unrolling the shutter: CNN to correct motion distortions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2291–2299.
[15] E. Ringaby and P.-E. Forssén, "Efficient video rectification and stabilisation for cell-phones," *International Journal of Computer Vision*, vol. 96, no. 3, pp. 335–352, 2012.
[16] E. Naor, I. Antebi, S. Bagon, and M. Irani, "Combining internal and external constraints for unrolling shutter in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 119–134.
[17] D. Qu, B. Liao, H. Zhang, O. Ait-Aider, and Y. Lao, "Fast rolling shutter correction in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 11 778–11 795, 2023.
[18] Z. Zhong, Y. Zheng, and I. Sato, "Towards rolling shutter correction and deblurring in dynamic scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9219–9228.
[19] B. Fan, Y. Mao, Y. Dai, Z. Wan, and Q. Liu, "Joint appearance and motion learning for efficient rolling shutter correction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 5671–5681.

[20] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. 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 (CVPR)*, 2018, pp. 9000–9008.

[21] S. Niklaus and F. Liu, "Softmax splatting for video frame interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5437–5446.

[22] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, "Real-time intermediate flow estimation for video frame interpolation," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 624–642.

[23] B. Fan, Y. Dai, and H. Li, "Rolling shutter inversion: Bring rolling shutter images to high framerate global shutter video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 6214–6230, 2022.

[24] D. Qu, Y. Lao, Z. Wang, D. Wang, B. Zhao, and X. Li, "Towards nonlinear-motion-aware and occlusion-robust rolling shutter correction," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023, pp. 10 680–10 688.

[25] Z. Zhong, M. Cao, X. Sun, Z. Wu, Z. Zhou, Y. Zheng, S. Lin, and I. Sato, "Bringing rolling shutter images alive with dual reversed distortion," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 233–249.

[26] X. Zhou, P. Duan, Y. Ma, and B. Shi, "EvUnroll: Neuromorphic events based rolling shutter image correction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 775–17 784.

[27] W. Shang, D. Ren, C. Feng, X. Wang, L. Lei, and W. Zuo, "Self-supervised learning to bring dual reversed rolling shutter images alive," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023, pp. 13 086–13 094.

[28] F. A. Reda, D. Sun, A. Dundar, M. Shoeybi, G. Liu, K. J. Shih, A. Tao, J. Kautz, and B. Catanzaro, "Unsupervised video interpolation using cycle consistency," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 892–900.

[29] Y.-L. Liu, Y.-T. Liao, Y.-Y. Lin, and Y.-Y. Chuang, "Deep video frame interpolation using cyclic frame generation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019, pp. 8794–8802.

[30] Y. Zhang, H. Wang, H. Zhu, and Z. Chen, "Optical flow reusing for high-efficiency space-time video super resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 5, pp. 2116–2128, 2022.

[31] M. Hu, J. Xiao, L. Liao, Z. Wang, C.-W. Lin, M. Wang, and S. Satoh, "Capturing small, fast-moving objects: Frame interpolation via recurrent motion enhancement," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3390–3406, 2021.

[32] M. Park, H. G. Kim, S. Lee, and Y. M. Ro, "Robust video frame interpolation with exceptional motion map," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 754–764, 2020.

[33] D. Rüfenacht, R. Mathew, and D. Taubman, "Temporal frame interpolation with motion-divergence-guided occlusion handling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 293–307, 2018.

[34] J. Park, K. Ko, C. Lee, and C.-S. Kim, "BMBC: Bilateral motion estimation with bilateral cost volume for video interpolation," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 109–125.

[35] L. Kong, B. Jiang, D. Luo, W. Chu, X. Huang, Y. Tai, C. Wang, and J. Yang, "IFRNet: Intermediate feature refine network for efficient frame interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1969–1978.

[36] G. Choi, P. Heo, and H. Park, "Triple-frame-based bi-directional motion estimation for motion-compensated frame interpolation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1251–1258, 2018.

[37] S. Meyer, A. Djelouah, B. McWilliams, A. Sorkine-Hornung, M. Gross, and C. Schroers, "Phasenet for video frame interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 498–507.

[38] M. Choi, H. Kim, B. Han, N. Xu, and K. M. Lee, "Channel attention is all you need for video frame interpolation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020, pp. 10 663–10 671.

[39] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4463–4471.

[40] W. Shangguan, Y. Sun, W. Gan, and U. S. Kamilov, "Learning cross-video neural representations for high-quality frame interpolation," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 511–528.

[41] D. Sun, X. Yang, M.-Y. Liu, and J. 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 (CVPR)*, 2018, pp. 8934–8943.

[42] Z. Liu, Z. Li, W. Chen, X. Wu, and Z. Liu, "Unsupervised optical flow estimation for differently exposed images in ldr domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5332–5344, 2023.

[43] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 402–419.

[44] J. Park, C. Lee, and C.-S. Kim, "Asymmetric bilateral motion estimation for video frame interpolation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 14 539–14 548.

[45] X. Xu, L. Siyao, W. Sun, Q. Yin, and M.-H. Yang, "Quadratic video interpolation," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[46] Y. Liu, L. Xie, L. Siyao, W. Sun, Y. Qiao, and C. Dong, "Enhanced quadratic video interpolation," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 41–56.

[47] Z. Chi, R. Mohammadi Nasiri, Z. Liu, J. Lu, J. Tang, and K. N. Plataniotis, "All at once: Temporally adaptive multi-frame interpolation with advanced motion modeling," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 107–123.

[48] H. Sim, J. Oh, and M. Kim, "XVFI: Extreme video frame interpolation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 14 489–14 498.

[49] Y. Li, Y. Zhu, R. Li, X. Wang, Y. Luo, and Y. Shan, "Hybrid warping fusion for video frame interpolation," *International Journal of Computer Vision*, vol. 130, no. 12, pp. 2980–2993, 2022.

[50] S. Niklaus and F. Liu, "Context-aware synthesis for video frame interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1701–1710.

[51] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang, "Depth-aware video frame interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3703–3712.

[52] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, 2019.

[53] W. Bao, W.-S. Lai, X. Zhang, Z. Gao, and M.-H. Yang, "MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 933–948, 2021.

[54] L. Lu, R. Wu, H. Lin, J. Lu, and J. Jia, "Video frame interpolation with transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 3532–3542.

[55] J. Park, J. Kim, and C.-S. Kim, "BiFormer: Learning bilateral motion estimation via bilateral transformer for 4K video frame interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 1568–1577.

[56] Z. Wang, X. Ji, J.-B. Huang, S. Satoh, X. Zhou, and Y. Zheng, "Neural global shutter: Learn to restore video from a rolling shutter camera with global reset feature," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 794–17 803.

[57] X. Ji, Z. Wang, Z. Zhong, and Y. Zheng, "Rethinking video frame interpolation from shutter mode induced degradation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023, pp. 12 259–12 268.

[58] C. Albl, Z. Kukelova, V. Larsson, M. Polic, T. Pajdla, and K. Schindler, "From two rolling shutters to one global shutter," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2505–2513.

[59] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros, "Learning dense correspondence via 3D-guided cycle consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 117–126.

[60] X. Wang, A. Jabri, and A. A. Efros, "Learning correspondence from the cycle-consistency of time," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2566–2576.

[61] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "Temporal cycle-consistency learning," in *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1801–1810.
[62] Z. Lai and W. Xie, "Self-supervised video representation learning for correspondence flow," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2019, p. 299.
[63] H. Wu and X. Wang, "Contrastive learning of image representations with cross-video cycle-consistency," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 10 149–10 159.
[64] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2223–2232.
[65] H. Wu, X. Zhang, W. Xie, Y. Zhang, and Y. Wang, "Boost video frame interpolation via motion adaptation," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2023, pp. 179–181.
[66] W. He, K. You, Z. Qiao, X. Jia, Z. Zhang, W. Wang, H. Lu, Y. Wang, and J. Liao, "TimeReplayer: Unlocking the potential of event cameras for video interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 804–17 813.
[67] M. Meingast, C. Geyer, and S. Sastry, "Geometric models of rolling-shutter cameras," *arXiv preprint arXiv:cs/0503076*, 2005.
[68] L. Oth, P. Furgale, L. Kneip, and R. Siegwart, "Rolling shutter camera calibration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1360–1367.
[69] B. Zhuang, L.-F. Cheong, and G. Hee Lee, "Rolling-shutter-aware differential SfM and image rectification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 948–956.
[70] B. Zhuang and Q.-H. Tran, "Image stitching and rectification for hand-held cameras," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 243–260.
[71] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, "GMFlow: Learning optical flow via global matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 8121–8130.
[72] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 694–711.
[73] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of International Conference on Learning Representations*, 2015.
[74] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.
[75] P.-E. Forssén and E. Ringaby, "Rectifying rolling shutter video from hand-held devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 507–514.
[76] B. Fan, Y. Dai, and H. Li, "Learning bilateral cost volume for rolling shutter temporal super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3862–3879, 2024.

Bin Fan received the B.S. degree, M.E. degree, and Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2016, 2019, and 2023, respectively. He was selected to CVPR 2022 Doctoral Consortium (the only one among Chinese universities). He co-organized the CVPR 2023 and ACCV 2022 tutorials on the topic of rolling shutter cameras. His research interests include computer vision, computational photography, and deep learning, especially regarding the rolling shutter camera.

Ying Guo received the M.E. degree in electronic information from Northwestern Polytechnical University, Xi'an, China, in 2024. She is currently pursuing the Ph.D. degree in information and communication engineering with the School of Electronics and Information, Northwestern Polytechnical University. Her research interests include computer vision, image processing, and deep learning.

Yuchao Dai received the B.E. degree, M.E degree and Ph.D. degree all in signal and information processing from Northwestern Polytechnical University (NPU), Xi'an, China, in 2005, 2008 and 2012, respectively. He is currently a Professor with School of Electronics and Information at NPU. He was an ARC DECRA Fellow with the Research School of Engineering at the Australian National University, Canberra, Australia. He won the Best Paper Award in IEEE CVPR 2012, the Best Paper Award Nominee at IEEE CVPR 2020, the DSTO Best Fundamental Contribution to Image Processing Paper Prize at DICTA 2014, the Best Algorithm Prize in NRSFM Challenge at CVPR 2017, the Best Student Paper Prize at DICTA 2017, the Best Deep/Machine Learning Paper Prize at APSIPA ASC 2017. He served/serves as Area Chair in CVPR, ICCV, NeurIPS, ACM MM etc. He serves as Publicity Chair in ACCV 2022 and Distinguished Lecture at APSIPA. His research interests include structure from motion, multi-view geometry, low-level computer vision, deep learning, compressive sensing and optimization.

Chao Xu received the B.E. degree from Tsinghua University, Beijing, China, in 1988, the M.S. degree from the University of Science and Technology of China, Hefei, China, in 1991, and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 1997. From 1991 to 1994, he was an Assistant Professor with the University of Science and Technology of China. Since 1997, he has been with the School of Electronics Engineering and Computer Science (EECS), Peking University, Beijing, where he is currently a professor. His research interests are in image and video coding, processing, and understanding. He has authored or coauthored more than publications and five patents in these fields.

Boxin Shi received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the PhD degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Associate Professor (with tenure) and Research Professor at Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. His papers were awarded as Best Paper, Runners-Up at CVPR 2024, ICCP 2015 and selected as Best Paper candidate at ICCV 2015. He is an associate editor of TPAMI/IJCV and an area chair of CVPR/ICCV/ECCV. He is a senior member of IEEE.