Noname manuscript No.
(will be inserted by the editor)

Self-Supervised Shutter Unrolling with Events

Mingyuan Lin · Yangguang Wang · Xiang Zhang · Boxin Shi · Wen Yang · Chu He · Gui-song Xia · Lei Yu

Received: date / Accepted: date

Abstract Continuous-time Global Shutter Video Recovery (CGVR) faces a substantial challenge in recovering undistorted high frame-rate Global Shutter (GS) videos from distorted Rolling Shutter (RS) images. This problem is severely ill-posed due to the absence of temporal dynamic information within RS intra-frame scanlines and inter-frame exposures, particularly when prior knowledge about camera/object motions is unavailable. Commonly used artificial assumptions on scenes/motions and data-specific characteristics are prone to producing sub-optimal solutions in real-world scenarios. To address this challenge, we propose an event-based CGVR network within a self-supervised learning paradigm, *i.e.*, *SelfUnroll*, and leverage the extremely high temporal resolution of event cameras to provide accurate inter/intra-frame dynamic information. Specifically, an Event-based Inter/intra-frame Compensator (E-IC) is proposed to predict the per-pixel dynamic between arbitrary time intervals, including the temporal transition and spatial translation. Exploring connections in terms of RS-RS, RS-GS, and GS-RS, we explicitly formulate mutual constraints with the proposed E-IC, resulting in supervisions without ground-truth GS images. Extensive evaluations over synthetic and real datasets demonstrate that the proposed method achieves state-of-the-art methods and shows remarkable performance for event-based RS2GS inversion in real-world scenarios. The dataset and code are available at https://w3un.github.io/selfunroll/.

Keywords Rolling Shutter Correction · Event Camera · Self-supervised Learning

Mingyuan Lin [1†]
linmingyuan@whu.edu.cn

Yangguang Wang [2†]
wangyangguang@xiaomi.com

Xiang Zhang [3]
xiangz.ethz@gmail.com

Boxin Shi [4]
shiboxin@pku.edu.cn

Wen Yang [1]
yangwen@whu.edu.cn

Chu He [1]
chuhe@whu.edu.cn

Gui-song Xia [5]
guisong.xia@whu.edu.cn

✉ Lei Yu [5]
ly.wd@whu.edu.cn

[1] School of Electronic and Information, Wuhan University, China.

[2] Xiaomi Communications Co., Ltd., China.

[3] Department of Computer Science, ETH Zurich, Switzerland.

[4] School of Computer Science, Peking University, China.

[5] School of Artificial Intelligence, Wuhan University, China.

[†] These authors contributed equally to this work.

1 Introduction

The row-by-row exposure mechanism in the Rolling Shutter (RS) camera significantly reduces the data transfer rate, making it an affordable solution for high-speed imaging (Zhong et al, 2022; Choi et al, 2022; Sheinin et al, 2022). However, such an approach produces so-called RS effects, which appear as spatial distortions in dynamic scenes (*e.g.*, as wobble and skew, as shown in Fig. 1), especially when high-speed camera-to-object motions are involved (Baker et al, 2010). Although correcting RS effects is essential for practical

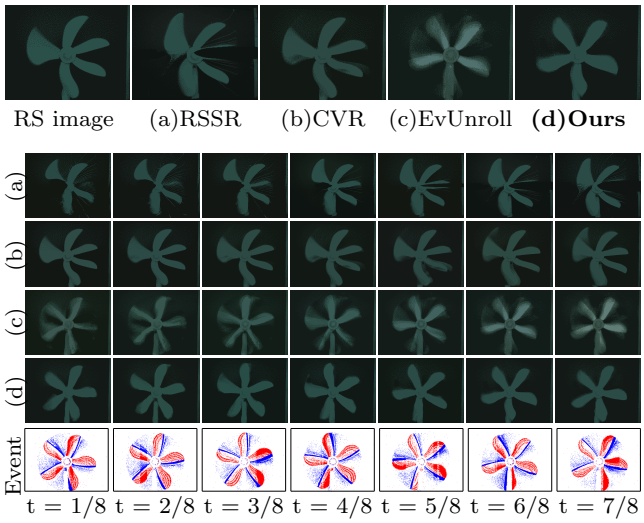


Fig. 1 A challenging example on real-world Gev-RS-Real dataset (Zhou et al, 2022) of unrolling RS images to GS video sequences. Frame-based methods, RSSR (Fan and Dai, 2021) and CVR (Fan et al, 2022), face difficulties in handling the typical non-linear motion. Event-based method EvUnroll (Zhou et al, 2022) suffers from "halo" artifacts and blurry details due to the distribution gap between synthetic training and real testing data. By contrast, our SelfUnroll achieves visually pleasant results by directly fitting real-world data distribution in a self-supervised manner.

applications, it is insufficient to completely capture the underlying continuous-time GS video. Alternatively, inverting the distorted RS observations to an undistorted continuous-time Global Shutter (GS) video can achieve the full Continuous-time Global shutter Video Recovery (CGVR), which, however, is inherently challenging due to the missing temporal dynamic information in both RS intra-frame scanlines and inter-frame exposures.

Recovering the GS frame from RS observations by removing the spatial distortion is highly ill-posed, especially when there is no supplementary information available on camera or object movements. Current methods tackle this issue by relying on artificial assumptions about scenes (Lao and Ait-Aider, 2018; Purkait and Zach, 2018; Rengarajan et al, 2016) and motions (Zhuang et al, 2019; Liu et al, 2020; Rengarajan et al, 2017) and thus fall short of accurate GS frame reconstruction. Moreover, these methods can only generate GS frames at a specific time due to the lack of *temporal dynamic information*. The CGVR task is further complicated by the need to recover undistorted GS frames for arbitrary timestamps, necessitating the recovery of lost intra-/inter-frame information in each moment. Recent methodologies address this requirement by assuming motion linearity on prior assumptions (Fan and Dai, 2021; Fan et al, 2022) or data-specific characteristics learned from specific datasets (Zhou et al, 2022). However, these artificial assumptions and data-specific characteristics do not always hold in real-world scenarios, leading to the significant performance degradation, as illustrated in Fig. 1 (a) and (b). Hence, the task of CGVR by inverting RS images to continuous-time GS videos in real-world scenarios still remains an ongoing and challenging research area.

In this paper, we propose to leverage inter-/intra-frame information with the aid of event cameras, a neuromorphic sensor that asynchronously emits events in response to the brightness change at an extremely high temporal resolution (Brandli et al, 2014; Gallego et al, 2020). While preliminary results have shown the feasibility of using events at the intra-frame time to correct RS distortions through EvUnroll (Zhou et al, 2022), the failure to utilize inter-frame information and the use of the linear motion assumption in the RS2GS flow module still falls short of achieving accurate CGVR. Furthermore, the model is trained solely on synthetic RS datasets, consisting of manually synthesized events and RS images, which limits its performance in real-world scenarios due to the "synthetic-to-real" domain gap caused by discrepancies in *data distribution* and *modality correspondence* (Fig. 1 (c)). The distribution gap between the synthetic and real datasets exists in either events or RS images due to the imperfection of physical cameras, *e.g.*, intrinsic thermal noise and event threshold variations (Brandli et al, 2014). In addition, the emission rate of events is confined to the read-out bandwidth (IniVation, 2020), resulting in temporal disorders of events that are difficult to simulate using event simulators (Rebecq et al, 2018; Delbruck et al, 2021). Furthermore, the per-pixel correspondence between the events and images is relatively simple and can be established in network training (Zhang et al, 2022; Pan et al, 2019; Zhang and Yu, 2022), but the learned correspondence in pre-trained models is fragile and can be disrupted by variations in the response functions of RS cameras in real-world scenes (Zhang and Yu, 2022).

Therefore, training with real events and RS images is essential to ensure optimal CGVR performance in real-world scenarios. However, collecting RS images with high frame-rate GS references for training purposes is sophisticated and requires the use of an additional expensive high-speed camera, such as Phantom VEO 640 ($ 67,500) used in Gev-RS (Zhou et al, 2022). And this highlights the need for cost-effective alternatives: *Can we learn CGVR from the real-world RS images without the ground-truth GS images?*

The answer is **YES**. We propose a novel self-supervised framework that can achieve accurate CGVR from distorted RS observations without the need for ground-truth GS images. The proposed framework can

perform RS to GS (RS2GS) conversion of arbitrary timestamps using real events and RS images, a significant advancement over existing methods. *To the best of our knowledge, this is the first attempt to apply self-supervised learning to event-based RS2GS conversion.* Specifically, we first develop an Event-based Inter/intra-frame Compensator (E-IC) to establish a unified spatial and temporal connection between RS and GS image domains via flow-based and synthesis-based techniques. Our E-IC takes flexible segments of events as inputs and thus can restore latent GS images at arbitrary timestamps without re-training, benefiting GS video extraction from a single RS frame. By exploring the cross-domain constraints from spatial and temporal perspectives, we train the E-IC network with real events and RS frames in a fully self-supervised manner and we call it **SelfUnroll-S**.

Taking into account the motion consistency and occlusion, we further propose a Motion and Occlusion Aware (MOA) module and extend **SelfUnroll-S** to **SelfUnroll-M**, which produces smoother GS videos even in the presence of occlusions. Finally, we construct a new dataset composed of real events and RS frames to facilitate the research of RS2GS conversion in real-world scenarios.

Overall, our contributions are threefold:

- We propose E-IC that allows for a unified and flexible transition between two images (RS2GS, GS2RS, and RS2RS) at arbitrary exposures, thereby enabling the continuous reconstruction of GS videos.
- We make the first attempt to approach self-supervised rolling shutter correction using an event camera, leveraging its effective perception of temporal dynamic information.
- Our proposed SelfUnroll outperforms state-of-the-art RS2GS approaches on publicly available benchmarks. Furthermore, our experiments demonstrate its proficiency on real-world datasets, including our newly constructed DRE dataset, which comprises pixel-aligned RS frames and DAVIS346-captured event streams. This dataset significantly aids in real-world event-based RS correction. The dataset and the corresponding code are readily accessible at https://w3un.github.io/selfunroll/.

We notice that a parallel investigation (Lu et al, 2023), which adopts a similar self-supervised framework for RS correction, was released approximately two months after the preprint release of our work[1]. Compared to (Lu et al, 2023), our self-supervised framework comprises not only the cycle and temporal consistency losses, *i.e.*, GS-RS and RS-RS in (Lu et al, 2023), but also the latent consistency loss, which leverages brightness and texture variations between rolling shutter (RS) images. All three constraints are derived for the E-IC modules to learn precise transformation between RS and GS images in the temporal and spatial domains using texture and motion information embedded in events. Our proposed SelfUnroll excels (Lu et al, 2023) in network design, learning strategy, and reconstruction accuracy.

[1] https://arxiv.org/abs/2304.06930

2 Related Work

Rolling Shutter Correction. Existing frame-based rolling shutter correction methods can be roughly categorized into *model-driven approaches* and *learning-based approaches*. For model-driven methods, Baker et al (2010) present a constant affine or translational distortion model to estimate the per-pixel motion vector from consecutive RS frames and correct the rolling shutter distortion. Rengarajan et al (2016) propose the rule "straight line must remain straight" to estimate camera motion by extracting curves. Purkait et al (2017) leverage geometric properties of the 3D scene to correct the distortion by estimating the orthogonal vanishing direction. Zhuang et al (2017) propose an RS-aware differential SfM algorithm, where the camera motion and dense depth map are utilized in an RS-aware warping for image rectification. Albl et al (2020) use a novel and effective dual-scanning (bottom-to-top and top-to-bottom) RS camera setup for RS correction. RS camera motion estimation problem (Grundmann et al, 2012; Liu et al, 2013; Lao and Ait-Aider, 2018; Purkait and Zach, 2018) can also be addressed with the aid of RANSAC (Fischler and Bolles, 1981).

Recently, learning-based approaches have been developed to achieve better RS2GS conversion performance. Rengarajan et al (2017) were the first to propose a CNN model to estimate camera motion parameters from a single RS image. Zhuang et al (2019) extend their previous work (Zhuang et al, 2017) and develop depth- and motion-aware models to predict dense depth maps and camera motions from a single RS image. Liu et al (2020) design a differentiable forward warping module that enables learning RS correction in an end-to-end manner. Fan et al (2021) leverage symmetric consistency constraint to aggregate the contextual cues. Zhong et al (2021) build the Joint Correction and Deblurring (JCD) network using a deformable attention module to simultaneously achieve RS correction and motion deblurring. However, most RS correction methods are designed to restore one GS image at a specific moment, and thus fail to extract and leverage the continuous-time GS video.

Video Frame Interpolation. Existing Video Frame Interpolation (VFI) approaches (Bao et al, 2019a,b; Liu et al, 2017) can be categorized into flow-based and kernel-based approaches. Flow-based approaches generally predict intermediate images using bidirectional optical flow (Jiang et al, 2018; Sun et al, 2018). Nevertheless, most assume uniform motion and linear optical flow between consecutive frames (Reda et al, 2019; Bao et al, 2019b), which may be violated in real scenes with complex and nonlinear motions. Kernel-based methods usually model the frame interpolation as the local convolution with reference frames (Niklaus et al, 2017a,b), which is more robust to brightness changes, but the scalability of kernel-based approaches is often limited by the fixed sizes of convolution kernels. Although VFI approaches are able to extract continuous-time GS videos, existing methods generally assume GS references and cannot be directly applied to RS inputs.

Event-based RS Correction and VFI. Event cameras are neuromorphic sensors that report asynchronous event streams in response to brightness changes (Lichtsteiner et al, 2008), which poses a paradigm shift in visual perception and enables almost continuous observation of dynamic scenes. Due to the extremely low latency, events implicitly encode inter/intra-frame information in terms of motions and textures (Gallego et al, 2020; Wang et al, 2020; Lin et al, 2020; Xu et al, 2021), which benefits both VFI (Tulyakov et al, 2021, 2022; He et al, 2022; Zhang and Yu, 2022) and RS correction (Zhou et al, 2022).

For VFI tasks, Timelens (Tulyakov et al, 2021) and Timelens++ (Tulyakov et al, 2022) are pioneer works that marry the advantages of warping-based and synthesis-based interpolation approaches, which can handle illumination changes and the sudden appearance of new objects between reference frames. Inspired by (Zhu et al, 2017), He et al (2022) design an unsupervised learning framework for video interpolation with event streams using cycle consistency. Zhang and Yu (2022) jointly solve the deblurring and interpolation problem by a Learnable Double Integral (LDI) network, which can generate high frame-rate sharp videos from consecutive blurry inputs. For RS correction, EvUnroll (Zhou et al, 2022) is the first attempt to recover GS frames during the intra-frame time, which partially achieves CGVR. EvShutter (Erbach et al, 2023) achieves unconstrained joint RS correction and deblurring without constant speed motion assumptions. However, the aforementioned event-based approaches focus on the intra-frame GS frame recovery and do not utilize the inter-frame events as the input. Hence, they cannot fully achieve CGVR from RS images.

CGVR from Rolling Shutter Images. CGVR from rolling shutter images, which is capable of generating the undistorted GS frames at any given timestamp combines RS correction and VFI, and few works have considered this challenging situation. Fan and Dai (2021) were the first to use a network called RSSR to extract a latent GS video sequence from two consecutive RS images. Under the assumption of a constant velocity, they convert the predicted optical flow to RS undistortion flow corresponding to scanlines. Furthermore, a context-aware architecture CVR (Fan et al, 2022) is proposed by leveraging a contextual aggregation procedure to alleviate the holes and artifacts caused by occlusions. Zhong et al (2022) propose IFED to merge the symmetric information of dual reversed rolling shutter distortion images and reconstruct GS video sequences. However, due to the limited inter-/intra-frame information, the aforementioned methods often struggle to restore accurate GS results and extract continuous-time GS video in real-world scenarios with complex motions.

Overall, the existing CGVR methods encounter challenges in achieving accurate GS results, particularly when dealing with complex non-linear motion. Besides, most methods are trained over synthetic datasets and often suffer from performance degradation in real-world scenarios due to the "synthetic-to-real" gap. To address these issues, we propose the E-IC to recover high frame-rate GS videos from RS frames and events, and develop a fully self-supervised method to fit real-world data distribution directly.

3 Method

We first formulate our task in Sec. 3.1 and then design the Event-based Inter/intra-frame Compensator (E-IC) in Sec. 3.2 based on a unified transition between pair of images (*i.e.*, RS2GS, GS2RS, and RS2RS). In Sec. 3.3, we introduce SelfUnroll-S built upon E-IC and present a self-supervised learning framework by utilizing the consistencies between RS and GS domains. We further develop a Motion and Occlusion Aware (MOA) module to handle occlusions by exploiting temporal information, which extends SelfUnroll-S to SelfUnroll-M in Sec. 3.4. Finally, we present the network details and event representation in Sec. 3.5.

3.1 Problem Formulation

RS and GS. Generally, we can define a continuous video $\mathcal{I} : \mathbb{R} \times \mathbb{Z}^2 \rightarrow \mathbb{R}_+$, and $\mathcal{I}(t, \mathbf{x})$ parameterized with time t and pixel location $\mathbf{x} \triangleq (x, y)$. The key distinction between GS and RS lies in the exposure mechanism. A

GS image I^G is globally exposed for all pixels, whereas an RS image I^R is exposed line-by-line. To formalize this, we introduce two location-aware time-shifting operators, $\mathcal{T}^R(\mathbf{x})$ and $\mathcal{T}^G(\mathbf{x})$, to define RS and GS images.

A GS image I^G is globally exposed for all pixels at t_0. Thus, setting $\mathcal{T}^G_{t_0}(\mathbf{x}) \triangleq t_0$, we have:

$$I^G_{t_0}(\mathbf{x}) = \mathcal{I}(\mathcal{T}^G_{t_0}(\mathbf{x}), \mathbf{x}). \tag{1}$$

An RS image I^R is exposed row-by-row, with the first row exposed at t_0:

$$I^R_{t_0}(\mathbf{x}) = \mathcal{I}(\mathcal{T}^R_{t_0}(\mathbf{x}), \mathbf{x}), \tag{2}$$

where $\mathcal{T}^R_{t_0}(\mathbf{x}) = t_0 + y\frac{|T|}{H}$, H is the image height, and T is the exposure time interval.

RS to GS (RS2GS). The RS2GS transition denotes the transformation from the RS domain to the GS domain, aimed at correcting distortions induced by the row-by-row exposure mechanism. This transition is mathematically expressed as:

$$I^G_{t_s} = \text{RS2GS}\left(I^R_{t_0}, t_s\right). \tag{3}$$

Drawing on the relationships defined in Eqs. (1) and (2), we can recast Eq. (3) to signify the transition between two exposure modes defined by $\mathcal{T}^R(\mathbf{x})$ and $\mathcal{T}^G(\mathbf{x})$, *i.e.*,

$$\mathcal{I}(\mathcal{T}^G_{t_s}(\mathbf{x}), \mathbf{x}) = \text{RS2GS}\left(\mathcal{I}(\mathcal{T}^R_{t_0}(\mathbf{x}), \mathbf{x}), t_s\right). \tag{4}$$

However, the transition is severely ill-posed because of the missing inter/intra-frame information.

RS2GS with events. Event cameras report asynchronous events whenever the brightness exceeds the event threshold $\eta > 0$ in the logarithmic domain (Brandli et al, 2014), *i.e.*,

$$\log(\mathcal{I}(t, \mathbf{x})) - \log(\mathcal{I}(\tau, \mathbf{x})) = p \cdot \eta, \tag{5}$$

where $\log(\mathcal{I}(t, \mathbf{x}))$ and $\log(\mathcal{I}(\tau, \mathbf{x}))$ denote the log-scale pixel brightness of position $\mathbf{x}$ at time t and τ, and $p \in \{+1, -1\}$ is the polarity indicating brightness increase $(+1)$ and decrease (-1). With the aid of events, the RS2GS transition Eq. (4) can be reformulated as:

$$\mathcal{I}(\mathcal{T}^G_{t_s}(\mathbf{x}), \mathbf{x}) = \text{RS2GS}\left(\mathcal{I}(\mathcal{T}^R_{t_0}(\mathbf{x}), \mathbf{x}), \mathcal{E}_{[\mathcal{T}^R_{t_0}, \mathcal{T}^G_{t_s}]}\right), \tag{6}$$

where $\mathcal{E}_{[\mathcal{T}^R_{t_0}, \mathcal{T}^G_{t_s}]}$ denotes the set of events triggered during $[\mathcal{T}^R_{t_0}, \mathcal{T}^G_{t_s}]$.

Existing RS2GS methods face two limitations: **a) Lack of dynamic information.** Frame-based methods focus primarily on restoring a GS image at the specific time (Liu et al, 2020; Zhong et al, 2021). However, the lack of dynamic information prevents these methods from restoring continuous-time GS frames for arbitrary timestamps. **b) Data inconsistency.** Previous works (Liu et al, 2020; Zhou et al, 2022) employ well-labeled synthetic datasets for supervision, which often leads to the performance drop in real scenes due to the "synthetic-to-real" data inconsistency.

Instead of solely focusing on restoring the GS image at a single timestamp, *e.g.*, $t_s = t_0 + \frac{T}{2}$, our work is dedicated to CGVR, with the goal of restoring the latent GS video $\mathcal{I}$ at any given timestamp $t_s \in \mathbb{R}$ with the assistance of events. Furthermore, to mitigate the sub-optimal solutions caused by the data distribution gap, we propose a self-supervised framework to adapt to real-world data distributions without the need for collecting high framerate videos $\mathcal{I}$ using an expensive high-speed camera.

3.2 Event-based Inter/intra-frame Compensator

In this section, we introduce a versatile transitional function named Event-based Inter/intra-frame Compensator (E-IC), which is not confined to RS2GS but enables flexible transitions between two images within three modes: RS2GS, RS2RS, and GS2RS. Given two different images I_s and I_d (either in RS or in GS) with time shifting operators $\mathcal{T}_s(\mathbf{x})$ and $\mathcal{T}_d(\mathbf{x})$, the E-IC can achieve the transition between I_s and I_d as:

$$\mathcal{I}(\mathcal{T}_d(\mathbf{x}), \mathbf{x}) = \text{E-IC}\left(\mathcal{I}(\mathcal{T}_s(\mathbf{x}), \mathbf{x}), \mathcal{E}_{[\mathcal{T}_s, \mathcal{T}_d]}\right). \tag{7}$$

By combing the Eqs. (1), (2) and (7) and setting $\mathcal{T}_s = \mathcal{T}^R_s$ and $\mathcal{T}_d = \mathcal{T}^G_d$, we can get:

$$\textbf{RS2GS:} \quad I^G_d = \text{E-IC}(I^R_s, \mathcal{E}_{[\mathcal{T}^R_s, \mathcal{T}^G_d]}). \tag{8}$$

By configuring different operators for $\mathcal{T}_s$ and $\mathcal{T}_d$, we can also formulate the GS2RS and RS2RS transitions:

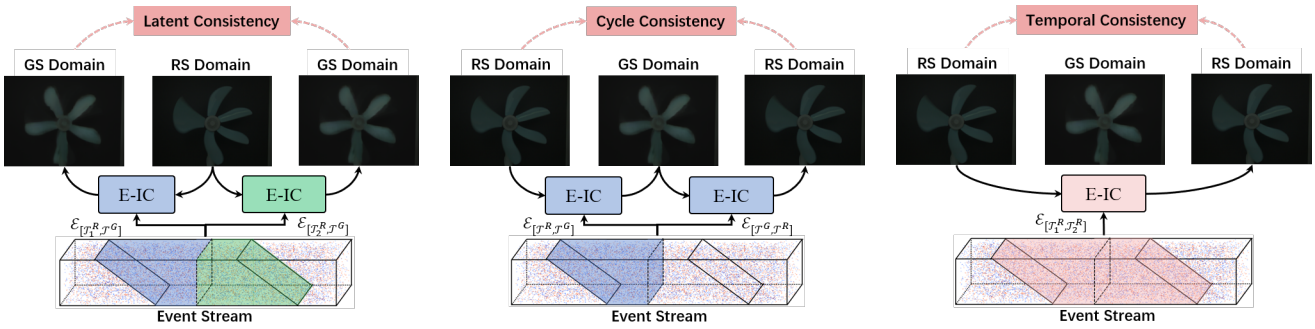
$$\begin{aligned} &\textbf{GS2RS:} \quad I^R_d = \text{E-IC}(I^G_s, \mathcal{E}_{[\mathcal{T}^G_s, \mathcal{T}^R_d]}), \\ &\textbf{RS2RS:} \quad I^R_d = \text{E-IC}(I^R_s, \mathcal{E}_{[\mathcal{T}^R_s, \mathcal{T}^R_d]}). \end{aligned} \tag{9}$$

As shown in Fig. 2, E-ICs can achieve flexible image transitions between RS and GS domains, *i.e.*, RS2GS (Fig. 2 (a)), GS2RS (Fig. 2 (b)), and RS2RS (Fig. 2 (c)), and further help us form a self-supervised learning framework that we will discuss in Sec. 3.3.1.

Inspired by the popular "warp and synthesis" approach from event-based video interpolation methods (Tulyakov et al, 2021), we define the E-IC in both temporal and spatial dimensions.

Temporal Transition. For a given coordinate $\mathbf{x}$, its temporal intensity transition can be formulated by integrating events over time,

$$\mathcal{I}(\mathcal{T}_d, \mathbf{x}) = \mathcal{I}(\mathcal{T}_s, \mathbf{x}) \cdot \tilde{\mathcal{I}}\left(\mathcal{E}_{[\mathcal{T}_s, \mathcal{T}_d]}, \mathbf{x}\right), \tag{10}$$



(a) Latent Consistency. (b) Cycle Consistency. (c) Temporal Consistency.

Fig. 2 Illustration of the proposed self-supervised learning framework based on the domain transformation function E-IC. (a) Latent consistency with RS2GS transition. (b) Cycle consistency with cascaded RS2GS and GS2RS transitions. (c) Temporal consistency with RS2RS transition.

with $\tilde{\mathcal{I}}$ denoting the brightness change between $\mathcal{T}_s, \mathcal{T}_d$,

$$\tilde{\mathcal{I}}\left(\mathcal{E}_{[\mathcal{T}_s,\mathcal{T}_d]}, \mathbf{x}\right) \triangleq \exp\left(\eta \int_{t=\mathcal{T}_s}^{\mathcal{T}_d} e(t,\mathbf{x})dt\right), \tag{11}$$

where $e(t,\mathbf{x}) \triangleq \sum_i p_i \delta(t-t_i)\delta(\mathbf{x}-\mathbf{x}_i)$ is the continuous representation of events and $\delta(\cdot)$ denoting the Dirac function.

Spatial Translation. Besides, benefiting from the low latency of event cameras, events naturally encode the motion information of dynamic scenes. Thus, given the source image, it is also feasible to achieve pixel translation in the spatial domain by utilizing the motion embedded in events, *i.e.*,

$$\mathcal{I}(\mathcal{T}_d, \mathbf{x}) = \mathcal{I}(\mathcal{T}_s, \mathbf{x} + \mathcal{F}(\mathcal{E}_{[\mathcal{T}_s,\mathcal{T}_d]})), \tag{12}$$

where $\mathcal{F}(\mathcal{E}_{[\mathcal{T}_s,\mathcal{T}_d]})$ denotes the optical flow estimated from events $\mathcal{E}_{[\mathcal{T}_s,\mathcal{T}_d]}$ triggered in the interval $[\mathcal{T}_s, \mathcal{T}_d]$.

However, directly computing the target image by Eq. (10) or (12) often suffers from the instability of the event threshold and the disturbance from noisy events, and thus we propose to employ deep learning-based methods for better pixel transition performance. In our approach, multiple Event-based Inter/intra-frame Compensators (E-ICs) are designed to take advantage of spatial and temporal compensations. We use two separate networks in our E-IC, *i.e.*, E-IC$_T$ to approximate Eq. (10) and E-IC$_S$ to approximate Eq. (12), and then fuse the results by a fusion module, which is denoted by E-IC$_{S+T}$, to get an combined E-IC output. Finally, the proposed E-IC can achieve a flexible transition between two images (in RS or GS) at any exposure time.

3.3 SelfUnroll with a Single RS Frame

The E-IC described in Sec. 3.2 establishes a unified transformation between images in RS and GS domains based on events. We implement E-IC for temporal intensity transition with the residual dense network (Jin et al, 2019) and for spatial pixel translation based on UNet (Ronneberger et al, 2015), and fuse two E-ICs to achieve RS correction from both spatial and temporal perspectives. We refer to this network as SelfUnroll-S (SelfUnroll with single RS frame) as shown in Fig. 3 (a), and propose a self-supervised learning framework for training.

3.3.1 Self-supervised Learning Framework

Based on E-ICs that achieve flexible image transitions between RS and GS domains, we devise the following three constraints and form a self-supervised learning framework.

Latent Consistency. Given two consecutive RS images I_1^R and I_2^R captured with exposure time $\mathcal{T}_1^R, \mathcal{T}_2^R$, we define the target GS image with exposure time $\mathcal{T}^G$. Then with event segments $\mathcal{E}_{[\mathcal{T}_1^R,\mathcal{T}^G]}, \mathcal{E}_{[\mathcal{T}_2^R,\mathcal{T}^G]}$, one can invert two RS images to the same latent GS image, leading to the latent consistency $\mathcal{L}_{lc}$ in the GS domain (Fig. 2 (a)), *i.e.*,

$$\mathcal{L}_{lc} = \left\| \text{E-IC}_*(I_1^R, \mathcal{E}_{[\mathcal{T}_1^R,\mathcal{T}^G]}) - \text{E-IC}_*(I_2^R, \mathcal{E}_{[\mathcal{T}_2^R,\mathcal{T}^G]}) \right\|_1, \tag{13}$$

where E-IC$_*$ is any one of E-IC$_S$,E-IC$_T$, and E-IC$_{S+T}$.

Cycle Consistency. We can formulate a cycle consistency $\mathcal{L}_{cc}$ in the RS domain by conducting the RS2GS process followed by a GS2RS process (Fig. 2 (b)), *i.e.*,

$$\mathcal{L}_{cc} = \left\| \text{E-IC}_*\left(\text{E-IC}_*(I^R, \mathcal{E}_{[\mathcal{T}^R,\mathcal{T}^G]}), \mathcal{E}_{[\mathcal{T}^G,\mathcal{T}^R]}\right) - I^R \right\|_1. \tag{14}$$

Temporal Consistency. Using the events between two consecutive RS frames, we can establish a temporal

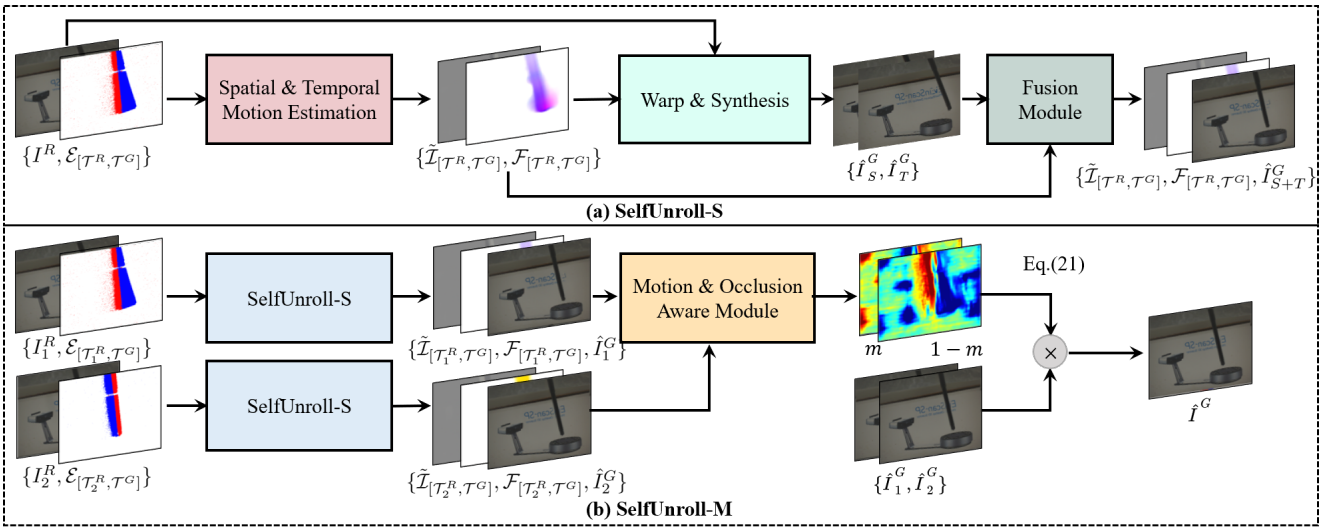


Fig. 3 The overall pipelines of SelfUnroll-S and SelfUnroll-M. **(a) SelfUnroll-S.** We establish the connection between the RS domain and the GS domain in both the spatial and temporal dimensions. And a fusion module is applied to take advantage of spatial and temporal compensation. **(b) SelfUnroll-M.** With two separate GS results $\hat{I}_1^G$ and $\hat{I}_2^G$, which are provided by SelfUnroll-S with two consecutive RS images I_1^R and I_2^R and share the same target exposure time $\mathcal{T}^G$, SelfUnroll-M can reconstruct the final result $\hat{I}^G$ with a motion and occlusion aware module.

consistency in the RS domain (Fig. 2 (c)), *i.e.*,

$$\begin{aligned}\mathcal{L}_{tc} =& \| \text{E-IC}_*(I_1^R, \mathcal{E}_{[\mathcal{T}_1^R, \mathcal{T}_2^R]}) - I_2^R \|_1 \\ &+ \| \text{E-IC}_*(I_2^R, \mathcal{E}_{[\mathcal{T}_2^R, \mathcal{T}_1^R]}) - I_1^R \|_1.\end{aligned} \tag{15}$$

In the self-supervised learning framework, $\mathcal{L}_{lc}$ supervises the structure of reconstruction by constraining the same latent GS image restored from different RS inputs. Using RS2GS and GS2RS conversions, $\mathcal{L}_{cc}$ ensures the stable brightness for image transitions between the RS and GS domains. Lastly, drawing on the widely-used principles of photometric consistency in video interpolation and optical flow estimation (Reda et al, 2019; Zhu et al, 2018), $\mathcal{L}_{tc}$ leverages information from adjacent RS frames, providing robust supervision for learning inter/intra-frame relationships from events.

3.3.2 Optimization

In addition to the three constraints described in Sec. 3.3.1, we also employ the Total Variation (TV) loss to smooth the flow map predicted in E-IC$_S$. Finally, the total self-supervisions can be summarized as follows,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{lc} + \lambda_2 \mathcal{L}_{cc} + \lambda_3 \mathcal{L}_{tc} + \lambda_4 \mathcal{L}_{tv}, \tag{16}$$

with λ_1, λ_2, λ_3, and λ_4 are balancing parameters and set as $\{1, 1, 1, 0.01\}$ for network training.

All E-ICs in our SelfUnroll network need to be trained with the above self-supervision consistencies, including E-IC$_S$, E-IC$_T$, and E-IC$_{S+T}$. And the whole network is trained in an end-to-end manner.

3.4 SelfUnroll with Multiple RS Frames

By exploiting the proposed E-ICs, our SelfUnroll-S can extract GS images at arbitrary timestamps from a single RS frame. However, real-world disturbances, *e.g.*, *foreground occlusions* and *noisy events*, might pose challenges to SelfUnroll-S. On the one hand, foreground occlusions often violate the brightness constancy assumption in E-IC$_S$. Although E-IC$_T$ can predict new objects in the scene, it tends to produce distorted colors due to the gap between monochrome events and color RS images. On the other hand, real-world events are noisy due to the non-ideality of event cameras, and the accumulation of noisy events often leads to performance degradation when inferring the GS images far from RS inputs. We develop a Motion and Occlusion Aware (MOA) module to address the above limitations by leveraging two consecutive RS images and events between them. Then we extend SelfUnroll-S to SelfUnroll-M to produce smoother and more accurate GS results.

3.4.1 Motion and Occlusion Aware Module

The proposed MOA module aims to utilize the temporal motion information from multiple inputs and improve the robustness of SelfUnroll against foreground occlusions and noisy events. Specifically, given two consecutive RS images I_1^R, I_2^R captured during the exposure time $\mathcal{T}_1^R, \mathcal{T}_2^R$ and the target GS image I^{GS} with exposure time $\mathcal{T}^G$, we first employ SelfUnroll-S to restore two separate GS results $\hat{I}_1^G, \hat{I}_2^G$ from I_1^R, I_2^R, *i.e.*,

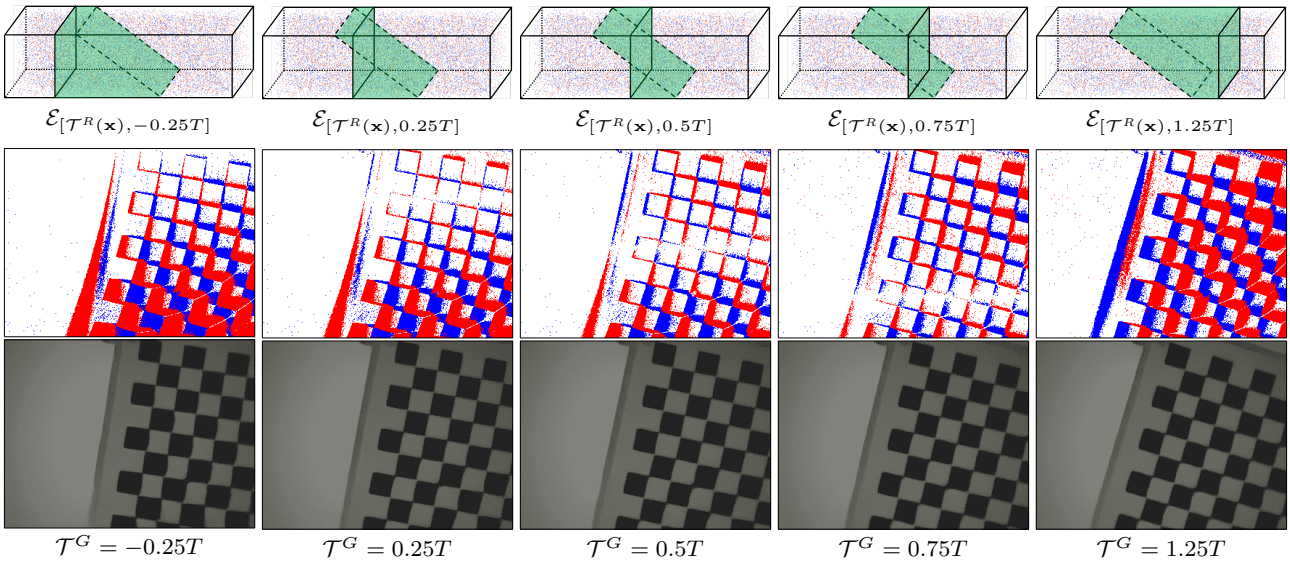


Fig. 4 Illustration of event segments $\mathcal{E}_{[\mathcal{T}^R,\mathcal{T}^G]}$ (1^{st} row) and corresponding event frames (2^{nd} row) for RS2GS inversions (3^{rd} row) of different latent GS timestamps $\mathcal{T}^G \in \{-0.25T, 0.25T, 0.5T, 0.75T, 1.25T\}$. Note that $\mathcal{T}^G = -0.25T$ or $1.25T$ represents the reconstructed GS frames not in the RS exposure time interval $[0, T]$.

$$\begin{aligned}\{\hat{I}_1^G, \mathcal{F}_1, \tilde{\mathcal{I}}_1\} &= \text{SelfUnroll-S}(I_1^R, \mathcal{E}_{[\mathcal{T}_1^R,\mathcal{T}^G]}),\\ \{\hat{I}_2^G, \mathcal{F}_2, \tilde{\mathcal{I}}_2\} &= \text{SelfUnroll-S}(I_2^R, \mathcal{E}_{[\mathcal{T}_2^R,\mathcal{T}^G]}),\end{aligned} \tag{17}$$

where the definition of $\mathcal{F}_1$, $\mathcal{F}_2$, $\tilde{\mathcal{I}}_1$, and $\tilde{\mathcal{I}}_2$ can be found in Eqs. (11) and (12), and the coordinate $\mathbf{x}$ is omitted for readability.

Then we feed the spatial and temporal transition information outputs from SelfUnroll-S and corresponding events to the MOA module to predict the confidence map m as:

$$m \triangleq \text{MOA}(\{\tilde{\mathcal{I}}_1, \mathcal{F}_1, \mathcal{E}_{[\mathcal{T}_1^R,\mathcal{T}^G]}\}, \{\tilde{\mathcal{I}}_2, \mathcal{F}_2, \mathcal{E}_{[\mathcal{T}_2^R,\mathcal{T}^G]}\}), \tag{18}$$

to predict the final reconstruction result $\hat{I}^G$ as:

$$\hat{I}^G = m\hat{I}_1^G + (1-m)\hat{I}_2^G. \tag{19}$$

As shown in Fig. 3 (b), SelfUnroll-M combines a SelfUnroll-S (or two weight-sharing SelfUnroll-S) and the MOA module. The SelfUnroll-M is fed with a pair of inputs, *i.e.*, two consecutive RS frames I_1^R, I_2^R and the events between the target GS frame and input RS frames $\mathcal{E}_{[\mathcal{T}_1^R,\mathcal{T}]}, \mathcal{E}_{[\mathcal{T}_2^R,\mathcal{T}]}$, and outputs the GS frame $\hat{I}^G$. For clarity, we summarize SelfUnroll-M as:

$$\hat{I}^G = \text{SelfUnroll-M}(I_1^R, I_2^R, \mathcal{E}_{[\mathcal{T}_1^R,\mathcal{T}^G]}, \mathcal{E}_{[\mathcal{T}_2^R,\mathcal{T}^G]}). \tag{20}$$

3.4.2 Optimization

We employ a two-stage training approach for SelfUnroll-M. Initially, we pre-train SelfUnroll-S using Eq. (16). Afterward, while keeping the weights of SelfUnroll-S fixed, we introduce a dual cycle consistency term $\mathcal{L}_{dcc}$ to train the MOA module self-supervisedly. Based on Eq. (20), we first employ SelfUnroll-M to restore an intermediate GS image $\hat{I}^G$ at the exposure time $\mathcal{T}^G$ from two RS frames I_1^{RS}, I_2^R and the corresponding events $\mathcal{E}_{[\mathcal{T}_1^R,\mathcal{T}^G]}, \mathcal{E}_{[\mathcal{T}_2^R,\mathcal{T}^G]}$. By exploiting the flexible conversion ability of SelfUnroll-M between RS and GS domains, we treat $\hat{I}^G$ as one reference image and estimate the original RS inputs,

$$\begin{aligned}\hat{I}_1^R &= \text{SelfUnroll-M}\left(\hat{I}^G, I_2^R, \mathcal{E}_{[\mathcal{T}^G,\mathcal{T}_1^R]}, \mathcal{E}_{[\mathcal{T}_2^R,\mathcal{T}_1^R]}\right),\\ \hat{I}_2^R &= \text{SelfUnroll-M}\left(\hat{I}^G, I_1^R, \mathcal{E}_{[\mathcal{T}^G,\mathcal{T}_2^R]}, \mathcal{E}_{[\mathcal{T}_1^R,\mathcal{T}_2^R]}\right).\end{aligned} \tag{21}$$

Then, one can formulate the dual cycle consistency loss,

$$\mathcal{L}_{dcc} = \left\|\hat{I}_1^R - I_1^R\right\|_1 + \left\|\hat{I}_2^R - I_2^R\right\|_1. \tag{22}$$

3.5 Network Details and Event Representation

Network Details. We adopt the residual dense network (Jin et al, 2019) as E-IC$_T$. The E-IC$_S$, fusion module, and MOA module are implemented based on UNet (Ronneberger et al, 2015). Note that all the modules in our network are trained from scratch.

Event Segments. As described in Sec. 3.2, our proposed E-ICs enable flexible RS2RS, RS2GS, and GS2RS transformations based on the corresponding event segment $\mathcal{E}_{[\mathcal{T}^R,\mathcal{T}^R]}$, $\mathcal{E}_{[\mathcal{T}^R,\mathcal{T}^G]}$, $\mathcal{E}_{[\mathcal{T}^G,\mathcal{T}^R]}$. Here we explain the principle of event segments using the RS2GS case. Since the RS image is exposed row by row during

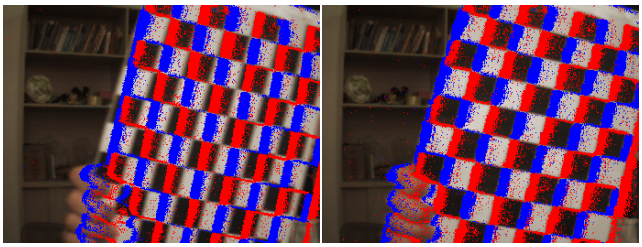
Table 1 Overview of datasets for event-based RS correction. Data*: including RS frames and events. †: the Gev-RS-Real dataset is built by two different event cameras with resolutions of 1280×720 and 346×260 respectively.

Datasets	Data*	Pixel Aligned	Resolution	Sequences
Fastec-RS (Liu et al, 2020)	synthetic	✓	640×480	75
Gev-RS (Zhou et al, 2022)	synthetic	✓	640×360	29
Gev-RS-Real (Zhou et al, 2022)	**real**	×	346×260†	16
DRE (Ours)	**real**	✓	346×260	**100**

exposure time interval $[0, T]$, each row has a row-specific exposure timestamp determined by the location-aware time shifting operator $\mathcal{T}^R(\mathbf{x})$. For the latent GS image, all pixels are exposed simultaneously at a single and fixed timestamp t. Thus, we define the event segment for the RS2GS transformation by $\mathcal{E}_{[\mathcal{T}^R,\mathcal{T}^G]} \triangleq \mathcal{E}_{[\mathcal{T}^R(\mathbf{x}),t]}$, corresponding to the events located at the green faded regions in Fig. 4. The event segment varies with respect to the latent GS timestamps and we give five illustrative examples including three GS images inside the RS exposure time ($t = 0.25, 0.5, 0.75T$ in Fig. 4) and two outside the RS exposure time ($t = -0.25, 1.25T$ in Fig. 4).

Event Representation. According to $\mathcal{E}_{[\mathcal{T}^R(\mathbf{x}),t]}$, events the different rows have different time intervals, therefore we stack events in a row-aware manner. For row y, we evenly divide N temporal bins ($N = 16$ in this work) between the latent GS timestamp t and the RS timestamp $\mathcal{T}^R(\mathbf{x}) = y\frac{|T|}{H}$ with H the image height and T the RS exposure time interval. To guarantee the performance of forward (*i.e.*, $\mathcal{T}^R(\mathbf{x}) < t$) and backward (*i.e.*, $\mathcal{T}^R(\mathbf{x}) > t$) conversion, we apply time flip and polarity reversal to events using the event pre-processing operator in (Zhang and Yu, 2022). We then accumulate the events inside each temporal bin and form a $2N \times 1 \times W$ tensor. The above operation is repeated for each row, and we finally concatenate them to form a $2N \times H \times W$ event tensor with $2, H, W$ indicating event polarity, image height, and width, respectively.

(a) Examples of scenes on the DRE dataset



(b) Before calibration (events in GS exposure)

(c) After calibration (events in RS exposure)

Fig. 5 Overview of the proposed DAVIS-RS-Event (DRE) dataset: (a) Indoor (top) and outdoor (bottom) samples captured under different camera motions and dynamic scenes; (b) misaligned events and RS image before line delay calibration; (c) aligned events and RS image after line delay calibration.

4 Experiment

4.1 Datasets

We evaluate the proposed algorithm on four datasets, including **two synthetic datasets**, *i.e.*, Fastec-RS (Liu et al, 2020) and Gev-RS-Sharp (Zhou et al, 2022), with simulated events and RS images, and **two real-world datasets**, *i.e.*, Gev-RS-Real (Zhou et al, 2022) and our proposed DAVIS-RS-Event (DRE), where DRE provides more sequences (100 *vs.* 16) with a richer diversity of scenes than Gev-RS-Real. We summarize the above datasets in Tab. 1.

Fastec-RS. The Fastec-RS (Liu et al, 2020) dataset captures multiple 2,400 FPS GS videos by a Fastec TS5 high-speed GS camera at the resolution of 640×480, and then synthesizes RS images. The corresponding event streams are synthesized with ESIM (Rebecq et al, 2018) using the high frame-rate GS video. We follow the same data splitting strategy as DSUN (Liu et al, 2020) and EvShutter (Erbach et al, 2023), *i.e.*, 54 sequences for training and 21 for testing.

Gev-RS-Sharp. The original Gev-RS (Zhou et al, 2022) dataset uses a high-speed Phantom VEO 640 camera to collect 5,700 FPS GS videos with the 640 × 360 image resolution and then synthesize event streams and motion blurred RS images. Since our SelfUnroll focus on the CGVR task and take sharp RS images as the input, we re-simulate the RS effects by the same method as Fastec-RS (Liu et al, 2020), where the sharp RS image is synthesized by sequentially copying a row of pixels from the high frame-rate GS images. With officially provided events, we formulate the Gev-RS-Sharp dataset has 29 paired sequences composed of 3,700 sharp GS-event-RS triplet clips. They are divided into the training and testing datasets by the same way as EvUnroll (Zhou et al, 2022).

Gev-RS-Real. The Gev-RS-Real dataset (Zhou et al, 2022) is built with real-world events and RS images cap-

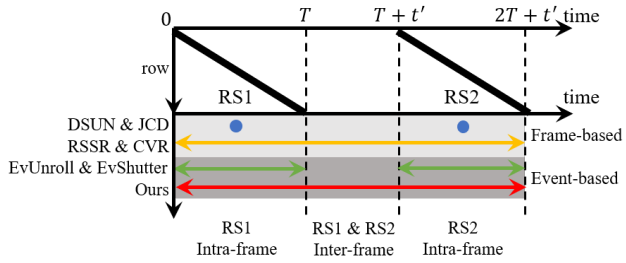


Fig. 6 Illustration of inter-frame and intra-frame time of RS video, where T represents the exposure length of RS frames and t' denotes the interval time between consecutive RS frames. The output settings of RS2GS approaches are demonstrated in the shaded regions. DSUN (Liu et al, 2020) and JCD (Zhong et al, 2021) only recover GS frame at a pre-defined timestamp, and EvUnroll (Zhou et al, 2022) and EvShutter(Erbach et al, 2023) are dedicated to restoring the intra-frame GS results. In contrast, RSSR (Fan and Dai, 2021), CVR (Fan et al, 2022), and our proposed SelfUnroll can achieve CGVR and restore GS frames in both intra-frame and inter-frame time.

tured by a hybrid camera system consisting of an RS camera and an event camera. The event streams are collected by two different event cameras, *i.e.*, DAVIS346 and PROPHESEE GEN4.0, for data diversity, and the ground-truth GS references are not available. Homography and a high-precision stopwatch are used for spatial and temporal synchronization. In total, Gev-RS-Real has 16 event-RS paired sequences, where we select 9 of them for training and the other 7 for testing.

DAVIS-RS-Event (DRE). Given that all existing event-based methods assume pixel alignment between frames and events, a condition not met in dynamic scenes for hybrid systems (Cho et al, 2023) like Gev-RS-Real, we have curated a large-scale, pixel-aligned real-world dataset, DAVIS-RS-Event (DRE). Utilizing a single DAVIS346 camera, we simultaneously captured real-world RS images and their corresponding event streams. Furthermore, we employed the Kalibr calibration tool (Oth et al, 2013) to determine the line delay (approximately $70\mu s$ in our calibration) between RS frames and events, as illustrated in Fig. 5. Our DRE dataset comprises more sequences (refer to Tab. 1) and showcases a richer diversity of scenes compared to the aforementioned datasets. This comprehensive dataset serves as a more thorough resource for RS2GS evaluation and training for unsupervised methods in real-world scenarios.

4.2 Implementation Details

Training. We use PyTorch (Paszke et al, 2019) to implement the proposed network with an NVIDIA GeForce RTX 3090 GPU. We randomly crop the RS images to 128×128 patches for training. Adam optimizer (Kingma and Ba, 2014) and SGDR scheduler (Loshchilov and Hutter, 2016) are employed with an initial learning rate set to 1×10^{-4}. Our models are trained in a two-stage manner: we first train the SelfUnroll-S using Eq. (16) for 100 epochs and then train SelfUnroll-M via Eq. (22) for another 100 epochs. Both SelfUnroll-S and SelfUnroll-M are trained with events and RS frames, and the ground-truth GS images are only used for performance evaluation.

Evaluation Metrics. For quantitative evaluation, we take the Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). Better reconstruction results are indicated by high PSNR and SSIM scores, and low LPIPS scores, *i.e.*, PSNR ↑, SSIM ↑, and LPIPS ↓.

4.3 Comparison on the Synthetic Datasets

Quantitative experiments are conducted on two synthetic datasets with GS references, *i.e.*, Gev-RS-Sharp (Zhou et al, 2022) and Fastec-RS (Liu et al, 2020). Our proposed SelfUnrolls are compared against state-of-the-art RS2GS methods, including four frame-based approaches, *i.e.*, DSUN (Fan et al, 2021), RSSR (Fan and Dai, 2021), JCD (Zhong et al, 2021) and CVR (Fan et al, 2022), as well as three event-based approach, EvUnroll (Zhou et al, 2022), EvShutter (Erbach et al, 2023), and NIRE (Zhang et al, 2024). We summarize the details of the abovementioned approaches in Fig. 6 and apply the same experimental settings to all methods. Metrics on the Fastec-RS dataset for RSSR and CVR are extracted from (Fan et al, 2022) and for the other methods are extracted from (Erbach et al, 2023). As the code for EvShutter is not publicly available, we only retrain DSUN, RSSR, JCD, CVR, and EvUnroll on the Gev-RS-Sharp dataset for a fair comparison. Metrics on the Gev-RS-Sharp dataset for NIRE are extracted from its paper. In the subsequent analysis, we divide CGVR into two subtasks, *i.e.*, the intra-frame and inter-frame reconstruction tasks as depicted in Fig. 6, and present comparisons in Secs. 4.3.1 and 4.3.2, respectively.

4.3.1 Comparison on Intra-frame Reconstruction

Quantitative comparisons on the Fastec-RS and Gev-RS-Sharp datasets are presented in Tabs. 2 and 3, demonstrating that event-based RS2GS approaches, *i.e.*, EvUnroll (Zhou et al, 2022), EvShutter (Erbach et al, 2023), NIRE (Zhang et al, 2024), and the proposed SelfUnrolls, *i.e.*, SelfUnroll-S and SelfUnroll-M, outperform frame-based methods by a large mar-

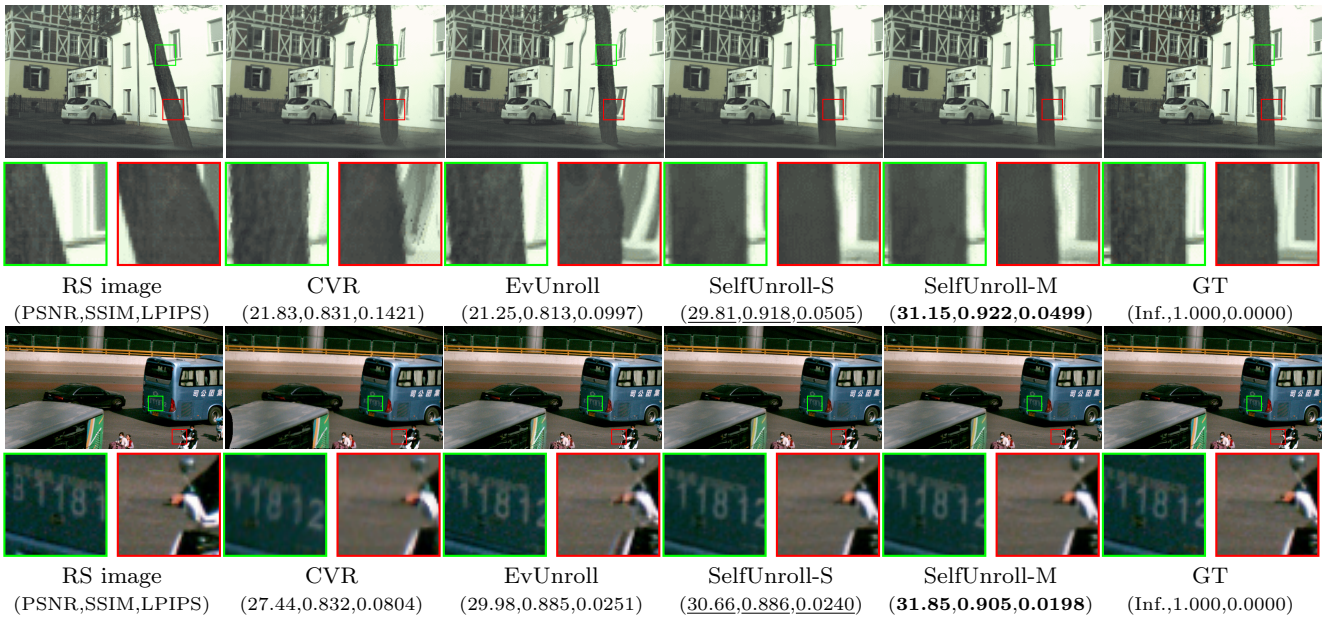


Fig. 7 Qualitative comparisons of single GS frame reconstruction on the Fastec-RS (top) and Gev-RS-Sharp (bottom) datasets. Details are zoomed in for a better view.

Table 2 Quantitative comparisons with respect to the single GS frame reconstruction on the Fastec-RS dataset. SSL represents self-supervised learning. **Bold** and underlined numbers represent the best and the second-best performance.

Methods	Event	SSL	PSNR↑	SSIM↑	LPIPS↓
DSUN (Liu et al, 2020)	×	×	26.52	0.79	0.122
JCD (Zhong et al, 2021)	×	×	24.84	0.78	0.107
RSSR (Fan and Dai, 2021)	×	×	21.23	0.78	0.142
CVR (Fan et al, 2022)	×	×	28.72	0.85	0.111
EvUnroll (Zhou et al, 2022)	✓	×	31.32	0.88	0.084
EvShutter (Erbach et al, 2023)	✓	×	32.41	**0.91**	**0.061**
SelfUnroll-S (Ours)	✓	✓	32.32	0.90	0.066
SelfUnroll-M (Ours)	✓	✓	**32.86**	**0.91**	0.064

Table 3 Quantitative comparisons of the proposed SelfUnrolls to the state-of-the-art methods on the Gev-RS-Sharp dataset. Given an RS frame with the exposure time $[0, T]$, all methods output 9 GS frames at timestamps $t \in \{0.1T, 0.2T, \ldots, 0.8T, 0.9T\}$ in the GS video sequence reconstruction task. In the single-frame reconstruction task, we evaluate the middle frame at the timestamp $t = 0.5T$ of the whole video. **Bold** and underlined numbers represent the best and the second-best performance. The symbol / denotes infeasible to reconstruct GS sequences.

Method	Single GS frame			GS video sequence		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DSUN (Liu et al, 2020)	23.95	0.839	0.0508	/	/	/
JCD (Zhong et al, 2021)	20.13	0.717	0.0718	/	/	/
RSSR (Fan and Dai, 2021)	22.21	0.749	0.1011	20.39	0.680	0.1195
CVR (Fan et al, 2022)	23.18	0.766	0.1022	23.52	0.771	0.1076
EvUnroll (Zhou et al, 2022)	31.29	0.914	0.0230	29.41	0.896	0.0383
NIRE (Zhang et al, 2024)	31.75	0.91	/	/	/	/
SelfUnroll-S (Ours)	32.26	0.926	0.0204	31.95	0.923	0.0231
SelfUnroll-M (Ours)	**32.62**	**0.932**	**0.0200**	**32.71**	**0.934**	**0.0194**

gin. This validates the assistance of the inter-/intra-frame information provided by events for RS correction. Among the event-based methods, our SelfUnrolls show favorable performance compared to EvUnroll and EvShutter. Note that in Tab. 3, the metrics of the re-trained JCD decrease compared to that reported in EvUnroll on the Gev-RS-Sharp dataset. This performance drop is because JCD was originally designed for the joint task of motion deblurring and RS correction while we retrain it on the sharp RS images.

The first row of Fig. 7 presents qualitative comparisons on the Fastec-RS dataset. It is observed that

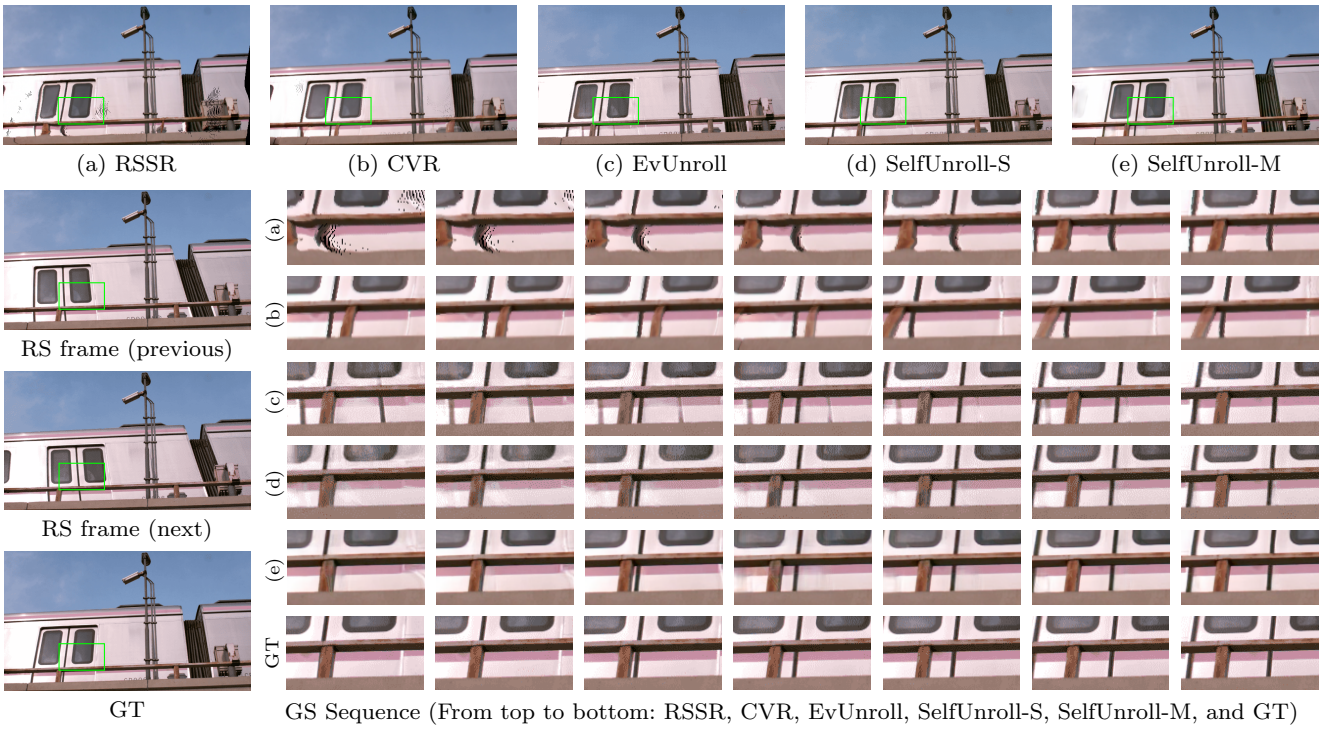


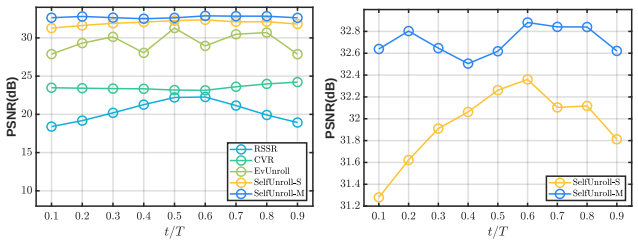
Fig. 8 Qualitative comparisons of GS video sequence restoration on the Gev-RS-Sharp dataset.

Table 4 Quantitative comparisons of the proposed SelfUnroll to the state-of-the-art NeRF-based RS correction methods on three sequences of the Gev-RS-Sharp dataset.

Method	24209_1_13			24209_1_30			24209_1_36		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
USB-NeRF (Li et al, 2024)	21.06	0.549	0.2356	15.85	0.494	0.5618	18.23	0.562	0.3050
RS-NeRF (Niu et al, 2025)	15.04	0.156	0.3051	11.51	0.189	0.7709	15.04	0.235	0.7324
SelfUnroll-S (Ours)	**30.63**	**0.960**	**0.0387**	**26.73**	**0.895**	**0.0493**	**33.60**	**0.983**	**0.0118**

both CVR and EvUnroll can restore the foreground tree but at the expense of distorting the background building. Regarding the reconstruction on the Gev-RS-Sharp dataset in the last row of Fig. 7, note that CVR fails to accurately estimate the underlying RS geometry, as demonstrated by the numbers on the car. On the other hand, although EvUnroll can correct the distortion, it also introduces artifacts on the edges of restored objects, as evident in the electric bicycle. In contrast, the proposed SelfUnrolls can rectify the edges of distorted objects and recover more realistic results, highlighting the effectiveness of the proposed algorithms.

Furthermore, we draw attention to the intra-frame GS video reconstruction task. Quantitative and qualitative comparisons are presented in Figs. 8 and 9 and Tab. 3. As shown in Fig. 8 (a) and (b), frame-based methods, such as RSSR and CVR, are unable to handle the dynamic scene with fast motion and fail to restore the shape. With the help of the motion and texture information that events provide, EvUnroll avoids the distortion of lines but suffers from severe artifacts



(a) All methods (b) SelfUnroll-S *vs.* -M

Fig. 9 Evaluation of Intra-frame GS video reconstruction at different target exposure times on the Gev-RS-Sharp dataset.

(Fig. 8 (c)). In contrast, SelfUnrolls can reconstruct intra-frame GS video with high stability and outperform other methods. Specifically, SelfUnroll-M can generate more reliable textures without color distortion, such as the train windows and the brown railing in Fig. 8 (e). Besides, the PSNR trend of GS video reconstruction in Fig. 9 (b) demonstrates that SelfUnroll-M can not only improve the single GS frame reconstruction quality, but also stabilize the performance of the GS video reconstruction. It can improve the reconstruc-

Table 5 Quantitative comparisons on the inter-frame GS video reconstruction task under the setting of 1- and 3-frame skips on the Gev-RS-Sharp dataset. Two RS frames with the exposure time $[0, T]$, $[T + t', 2T + t']$ (t' is the interval between two consecutive RS frames and is set to $\frac{2}{3}T$) act as the inputs. For the 1-frame skip task, all methods interpret 1 GS frame at the timestamp $t = T + 0.50t'$. For the 3-frame skip task, all methods interpret 3 GS frames at the timestamps $t \in \{T + 0.25t', T + 0.50t', T + 0.75t'\}$. **Bold** and underlined numbers represent the best and the second-best performance.

Methods	1 frame skip			3 frames skip		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
RSSR	16.02	0.532	0.2088	16.07	0.555	0.5333
CVR	23.84	0.760	0.1072	23.90	0.763	0.1064
EvUnroll+Timelens	19.17	0.726	0.1473	19.29	0.724	0.1480
EvUnroll+DAIN	19.29	0.688	0.1341	19.18	0.680	0.1479
SelfUnroll-S (Ours)	30.66	0.903	0.0323	30.68	0.903	0.0323
SelfUnroll-M (Ours)	**31.73**	**0.919**	**0.0315**	**31.77**	**0.920**	**0.0311**

Fig. 10 Single GS frame reconstruction of inter-frame time on the Gev-RS-Sharp dataset.

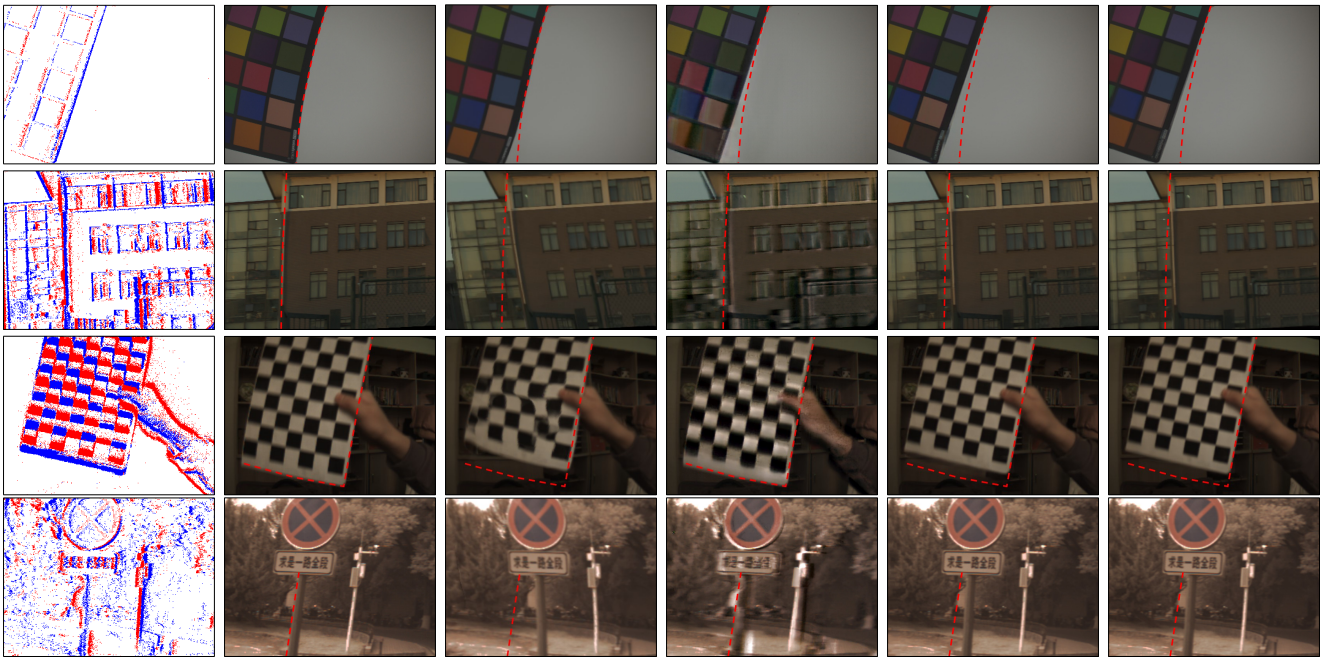
tion quality of GS images at edge target time by up to 1.3 dB on PSNR, shows the effectiveness of the MOA in the information fusion of consecutive frames.

Moreover, we conduct comparisons of our SelfUnroll to two NeRF-based RS correction methods, *i.e.*, USB-NeRF (Li et al, 2024) and RS-NeRF (Niu et al, 2025), on the Gev-RS-Sharp dataset. Since NeRF-based methods are based on two assumptions: *static scene (with moving camera)* and *known camera poses*. Therefore, we first exclude the sequences with moving objects from the testing set on the Gev-RS-Sharp dataset. Among the remaining sequences, we select the first 50 RS frames of each sequence for experiments, of which 40 frames are for training and 10 frames are for testing. Then, we estimate the camera poses using COLMAP (Schonberger and Frahm, 2016). Despite sequences lacking the diversity of camera views, *e.g.*, photographing distant buildings, COLMAP only recovers the poses of three sequences, *i.e.*, 24209_1_13, 24209_1_30, and 24209_1_36, which are used to initialize USB-NeRF and RS-NeRF. Quantitative results in Tab. 4 demonstrate that our SelfUnroll is more robust to real-world scenarios and image capture methods than NeRF-based methods.

4.3.2 Comparison on Inter-frame Reconstruction

We conduct the experiments on the inter-frame GS video reconstruction task of the proposed SelfUnrollto two frame-based methods, *i.e.*, RSSR (Fan and Dai, 2021) and CVR (Fan et al, 2022), and one evnet-bansed method, *i.e.*, EvUnroll (Zhou et al, 2022), on the Gev-RS-Sharp dataset. Since EvUnroll is not designed for the inter-frame GS video recovery task, we combine it with state-of-the-art video frame interpolation methods, *i.e.*, Timelens (Tulyakov et al, 2021) and DAIN (Bao et al, 2019a), denoting as EvUnroll+Timelens and EvUnroll+DAIN respectively.

Specifically, The RS frames are first corrected to the GS frames corresponding to the middle scanline by EvUnroll and then interpolated by Timelens or DAIN. As depicted in Tab. 5 and Fig. 10, SelfUnroll-S and SelfUnroll-M outperform the frame-based methods, RSSR and CVR, by a large margin, highlighting the importance of inter-frame scene dynamic information recorded by events. As for EvUnroll+Timelens and EvUnroll+DAIN, the cascade approaches experience a significant decline in performance as the RS2GS errors propagate to the frame interpolation stage. In contrast, our SelfUnrolls use E-ICs to transition between input and output images at any arbitrary moment rather than a specific time period, thereby maintaining high-quality RS2GS reconstruction during the inter-frame time. Considering the performance of SelfUnrolls, the advantage of SelfUnroll-M indicates the importance of consecutive frame information in inter-frame reconstruction.



(a) Event frame (b) RS frame (c) CVR (d) EvUnroll (e) SelfUnroll-S (f) SelfUnroll-M

Fig. 11 Qualitative results on the Gev-RS-Real dataset (top two rows) and DRE dataset (bottom two rows). All GS frame exposure times correspond to the first scanline of the RS image. The red dashed curves (with the same position and shape in (b)-(f)) indicate the distorted edges in the RS images and event frames represent the correct GS edges as the reference.

Table 6 Ablation study of E-ICs modules on the Gev-RS-Sharp dataset. **Bold** numbers denote the best performance.

Spatial	Temporal	PSNR↑	SSIM↑	LPIPS↓
✓		28.67	0.877	0.0311
	✓	31.67	0.916	0.0227
✓	✓	**32.26**	**0.926**	**0.0204**

Due to the different experimental settings between this paper and (Lu et al, 2023), whose codes are unavailable, we do not compare the proposed SelfUnroll with it. Nonetheless, it can be seen from the metrics provided by (Lu et al, 2023) that our SelfUnroll outperforms it by a large margin, *e.g.*, at least 6 dB, 0.13, and 0.06 improvements in terms of PSNR, SSIM, and LPIPS, on the Fastec-RS and Gev-RS-Sharp datasets.

4.4 Comparison on the Real-world Datasets

We also conduct experiments on real-world datasets, *i.e.*, Gev-RS-Real (Zhou et al, 2022) and DRE in Sec. 4.1. Qualitative comparisons are made to the state-of-the-art frame-based method CVR (Fan et al, 2022) and event-based method EvUnroll (Zhou et al, 2022). Fig. 11 illustrate comparisons on the single frame restoration, while the performance on the video reconstruction is shown in Figs. 1 and 12 respectively. We can observe that both CVR and EvUnroll are trained on synthetic datasets and thus suffer from degraded performance in real-world scenes due to the "synthetic-to-real" gap. Specifically, CVR produces inaccurate rectification and shape distortions due to violated motion assumptions, while EvUnroll yields obvious artifacts and blurred details due to the violation of learned modality correspondence in real-world scenes. In contrast, SelfUnrolls utilizes self-supervised learning

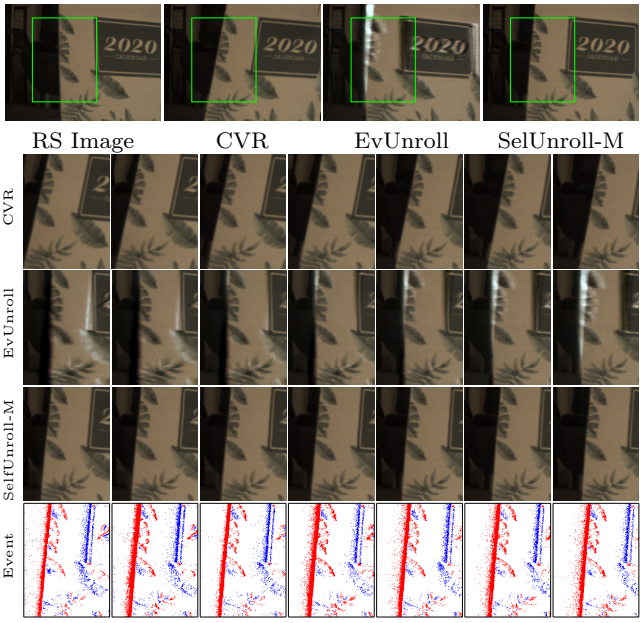


Fig. 12 Qualitative comparison on sequence restoration over the DRE dataset.

Fig. 13 Qualitative ablations of spatial and temporal E-ICs in SelfUnroll.

Table 7 Ablation study of supervisions on the Gev-RS-Sharp dataset. Case #0 represents the model without any supervision, and the metrics of Case #0 which are calculated between the input RS frame and the GS reference are given as a reference. **Bold** and underlined numbers represent the best and the second-best performance.

Case	$\mathcal{L}_{lc}$	$\mathcal{L}_{cc}$	$\mathcal{L}_{tc}$	$\mathcal{L}_{tv}$	SelfUnroll-S			SelfUnroll-M		
					PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
#0					19.01	0.673	0.0751	19.01	0.673	0.0751
#1	✓			✓	17.30	0.449	0.1110	17.45	0.439	0.1344
#2		✓		✓	19.01	0.670	0.0761	19.48	0.671	0.1102
#3			✓	✓	30.60	0.898	0.0277	30.85	0.905	0.0275
#4	✓	✓		✓	29.23	0.877	0.0944	29.47	0.882	0.0971
#5	✓		✓	✓	31.30	0.911	0.0243	31.56	0.915	0.0236
#6		✓	✓	✓	30.98	0.906	0.0248	31.37	0.914	0.0229
#7	✓	✓	✓		31.50	0.918	0.0227	32.15	0.925	0.0204
#8	✓	✓	✓	✓	**32.26**	**0.926**	**0.0204**	**32.62**	**0.932**	**0.0200**

on real datasets, effectively bridging the "synthetic-to-real" gap and producing visually satisfying results without texture and shape distortion. More results on the real-world datasets can be found at https://w3un.github.io/selfunroll/.

4.5 Ablation study

In this subsection, we perform a diverse set of ablation studies to investigate the contribution of each loss in our self-supervised learning framework and the importance of each module of the proposed network.

4.5.1 Importance of Spatial & Temporal E-ICs

The overall SelfUnroll-S network is composed of spatial and temporal E-ICs, *i.e.*, E-IC$_S$ and E-IC$_T$. The spatial connection E-IC$_S$ achieves RS2GS transformation by estimating per-pixel motions and thus has brightness consistency but may suffer from geometric distortions caused by motion errors from noisy events, as shown in Fig. 13. On the other hand, the temporal E-IC$_T$ module compensates for brightness change between RS and GS images and is thus robust to motion distortions but may suffer from unrealistic artifacts and chromatic aberrations, as shown in Fig. 13. Fusing spatial and temporal E-ICs gives the best quantitative and qualitative results, as shown in Fig. 13 and Tab. 6.

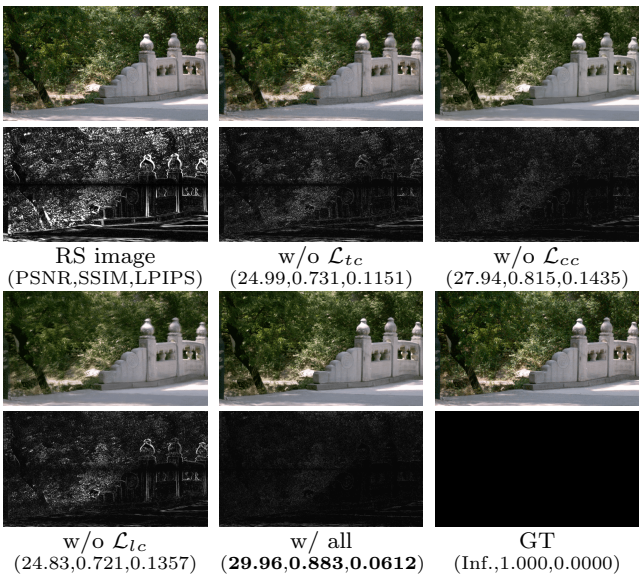


Fig. 14 Qualitative ablations of each supervision and their absolute differences to the ground-truth GS reference.

4.5.2 Importance of Combining Losses

According to Fig. 14, removing either $\mathcal{L}_{lc}$ or $\mathcal{L}_{tc}$ leads to RS distortions, suggesting that both components play a crucial role in correcting RS. Specifically, $\mathcal{L}_{tc}$ directly introduces the determined inputs, *i.e.*, two consecutive RS frames as the strong supervision to guide E-ICs to learn spatial and temporal transitions, which ulti-

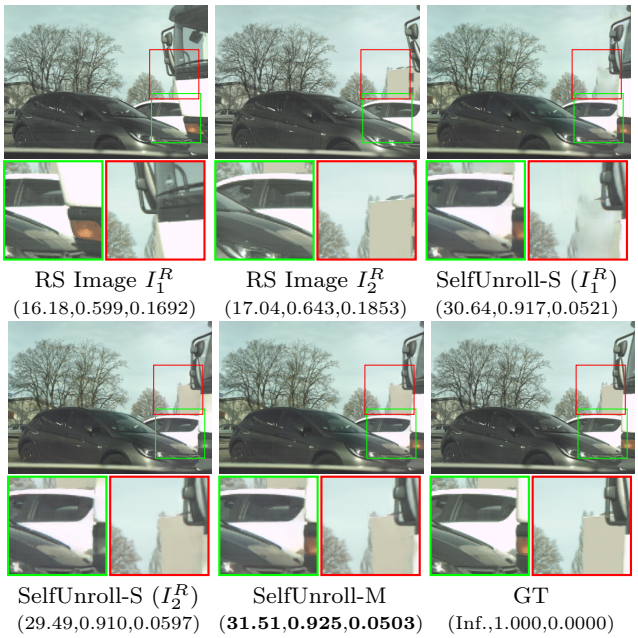


Fig. 15 Quantitative and qualitative ablations of MOA on the Fastec-RS dataset. I_1^R (previous) and I_2^R (next) are two consecutive RS frames.

mately benefits the RS2GS task. $\mathcal{L}_{lc}$ provides the only constraint in the GS domain and thus is also helpful in the RS2GS task. However, $\mathcal{L}_{lc}$ alone cannot guarantee brightness consistency as it does not impose a constraint between the outputs of E-ICs and the original images. Although $\mathcal{L}_{cc}$ does not directly help E-ICs produce GS images, it contributes to brightness consistency and improves the performance. Overall, Tab. 7 and Fig. 14 show that combining all three losses leads to the smallest absolute error, validating the necessity of supervision with $\mathcal{L}_{lc}$, $\mathcal{L}_{cc}$, and $\mathcal{L}_{tc}$ simultaneously.

4.5.3 Superiority of Multi-frame Fusion

The quantitative and qualitative analyses in Secs. 4.3 and 4.4 have demonstrated that SelfUnroll-M outperforms SelfUnroll-S due to the complementarity of the previous and next frame information. To study how the MOA module handles the information loss introduced by the occlusion, we compare the performance of SelfUnroll-S and SelfUnroll-M when facing the large occlusions, *e.g.*, the orange headlight and the distant building in Fig. 15. Using a single RS image (I_1^R/I_2^R) as input to SelfUnroll-S produces incorrect textures and colors in an attempt to recover regions occluded by foreground objects in moving scenes, *e.g.*, the disappearing building in SelfUnroll-S (I_1^R) and the gray headlight in SelfUnroll-S (I_2^R) in Fig. 15. Our MOA module can adaptively fuse the GS results generated by SelfUnroll-S (I_1^R/I_2^R) to alleviate the impact of occlusion.

4.6 Evaluation for RS Reconstruction

As detailed in Sec. 3.2, E-IC is designed not only for RS2GS transition but also encompasses GS2RS and RS2RS transitions. To comprehensively evaluate the reconstructive capabilities of E-IC on RS images, we conduct the GS2RS transition on the Gev-RS-Sharp dataset and the RS2RS transition on the DRE dataset, as depicted in Fig. 16. Furthermore, we explore the impact of different line delays (*e.g.,* $1\times$, $2\times$, and $-1\times$) to generate RS images with varying degrees of distortion.

As illustrated in Fig. 16, the RS reconstruction with $2\times t_d$ exhibits more distortion than $1\times t_d$, while the RS reconstruction with $-1\times t_d$ introduces distortion with a reversed direction. This experiment validates that our proposed E-IC can flexibly handle transitions to reconstruct both RS and GS images by utilizing different event segments as inputs.

4.7 Runtime and Performance Analysis

We further evaluate the complexity of our proposed approaches and other RS2GS methods by feeding RS images with the same spatial resolution 640×480 and executing them on the same NVIDIA GeForce RTX 3090 GPU. The comparison of runtime and performance is visualized in Fig. 17. Our SelfUnroll-S can infer GS prediction with the shortest time ($\approx$ 68 ms) while SelfUnroll-M performs with increased complexity since it requires an additional MOA module. SelfUnroll-M performs better than SelfUnroll-S while both achieve state-of-the-art RS2GS performance.

Regarding the model size, CVR has the most network parameters and performs best among frame-based RS2GS methods. Due to the utilization of events, the event-based methods have smaller model sizes but outperform CVR by a large margin. Among the three event-based methods, our proposed SelfUnroll-S gives higher GS reconstruction PSNR than EvUnroll even though they have comparable model sizes. The RS2GS performance can be further improved by SelfUnroll-M with increased model sizes and input frames. Overall, our proposed SelfUnroll-S/-M is highly effective and efficient in restoring the GS frame and facilitating the high temporal reconstruction of RS2GS.

5 Limitations and Future Works

Our proposed SelfUnroll method is based on the principle of estimating the spatial and temporal transformation between two sharp images. However, it is limited in its ability to handle motion blur, as shown in the

Fig. 16 Qualitative results of RS reconstruction. The top row illustrates the GS2RS transition using a GS image from the Gev-RS-Sharp dataset, while the bottom row displays the RS2RS transition utilizing RS images from the DRE dataset. t_d denotes the default line delay of RS images in the corresponding dataset, and $n \times t_d$ represents the line delay of reconstructed RS images. A higher value of n indicates a more pronounced RS effect. The minus sign "-" denotes reverse exposure order, where the exposure occurs from bottom to top (default RS exposure is from top to bottom in this work).

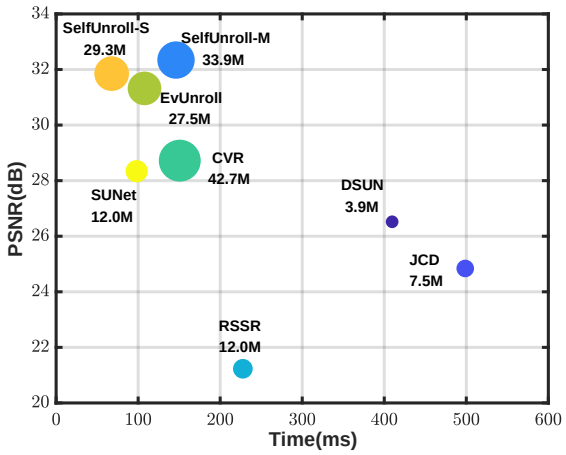
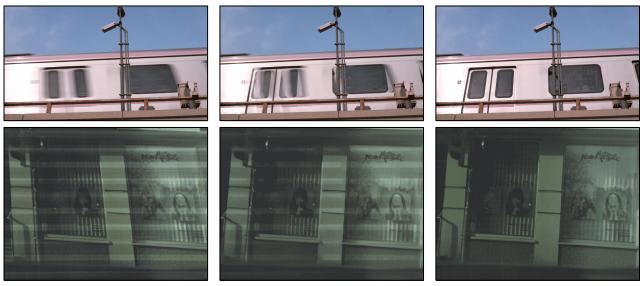


Fig. 17 PSNR *vs.* Inference time, size ∝ parameters on Fastec-RS dataset. The inference time is evaluated on reconstructing a GS image with a resolution of 640×480. The size of the blobs is proportional to the number of network parameters. The top left corner indicates the best performance.



(a) RS image (b) SelfUnroll-M (c) GT

Fig. 18 Failure cases: a blurry RS image showing a train with large motion (1^{st} row) and an RS image with light flicker effect (2^{nd} row).

first row of Fig. 18. Additionally, as illustrated in the second row of Fig. 18, our method is unable to mitigate the flicker effect (Lin et al, 2023) caused by the variation in brightness during RS exposure time. While we have successfully removed the distortion, the streaks caused by the flicker effects remain. We plan to address these issues in the future.

6 Conclusion

In this paper, we propose a novel event-based RS2GS method, SelfUnroll, to simultaneously achieve the RS image correction and the high temporal GS video reconstruction. Our method is based on the Event-based Inter/Intra-frame Compensator (E-IC), which establishes a unified spatial and temporal connection between two images with different exposures, thus improving the RS2GS transformation. Based on E-IC, we first propose the SelfUnroll-S network to restore GS images from a single RS image with the aid of events. Then, we extend SelfUnroll-S to SelfUnroll-M, where the Motion and Occlusion Aware (MOA) module is designed to tackle the occlusion problem by fusing the temporal information. We further propose a self-supervised learning framework for both SelfUnroll-S and SelfUnroll-M that allows for learning with real events and RS images, thereby reducing the cost of collecting ground-truth GS images and bridging the performance gap between synthetic and real datasets. To validate the effectiveness of our proposed methods, we develop a real-world RS image dataset that contains both events and RS frames

for various indoor and outdoor scenes. Our experimental results on synthetic and real datasets demonstrate the superiority of our proposed SelfUnroll methods over existing state-of-the-art methods.

Acknowledgements The research was partially supported by the National Natural Science Foundation of China under Grant 62271354.

Data Availability All the datasets used in this paper are publicly available. Fastec-RS is available at https://github.com/ethliup/DeepUnrollNet/. Gev-RS and Gev-RS-Real are available at https://github.com/zxyemo/EvUnroll/. Our DRE is available at https://w3un.github.io/selfunroll/.

References

Albl C, Kukelova Z, Larsson V, Polic M, Pajdla T, Schindler K (2020) From two rolling shutters to one global shutter. In: CVPR, pp 2505–2513

Baker S, Bennett E, Kang SB, Szeliski R (2010) Removing rolling shutter wobble. In: CVPR, pp 2392–2399

Bao W, Lai WS, Ma C, Zhang X, Gao Z, Yang MH (2019a) Depth-aware video frame interpolation. In: CVPR, pp 3703–3712

Bao W, Lai WS, Zhang X, Gao Z, Yang MH (2019b) Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. IEEE TPAMI 43(3):933–948

Brandli C, Berner R, Yang M, Liu SC, Delbruck T (2014) A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. IEEE J Solid-State Circuits 49(10):2333–2341

Cho H, Jeong Y, Kim T, Yoon KJ (2023) Non-coaxial event-guided motion deblurring with spatial alignment. In: ICCV, pp 12492–12503

Choi J, Wong CW, Su H, Wu M (2022) Analysis of ENF Signal Extraction From Videos Acquired by Rolling Shutters. TechRxiv

Delbruck T, Hu Y, He Z (2021) V2e: From video frames to realistic dvs event camera streams. In: CVPRW

Erbach J, Tulyakov S, Vitoria P, Bochicchio A, Li Y (2023) Evshutter: Transforming events for unconstrained rolling shutter correction. In: CVPR, pp 13904–13913

Fan B, Dai Y (2021) Inverting a rolling shutter camera: bring rolling shutter images to high framerate global shutter video. In: ICCV, pp 4228–4237

Fan B, Dai Y, He M (2021) Sunet: symmetric undistortion network for rolling shutter correction. In: ICCV, pp 4541–4550

Fan B, Dai Y, Zhang Z, Liu Q, He M (2022) Context-aware video reconstruction for rolling shutter cameras. In: CVPR, pp 17572–17582

Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24(6):381–395

Gallego G, Delbrück T, Orchard G, Bartolozzi C, Taba B, Censi A, Leutenegger S, Davison AJ, Conradt J, Daniilidis K, et al (2020) Event-based vision: A survey. IEEE TPAMI 44(1):154–180

Grundmann M, Kwatra V, Castro D, Essa I (2012) Calibration-free rolling shutter removal. In: 2012 IEEE International Conference on Computational Photography (ICCP), pp 1–8

He W, You K, Qiao Z, Jia X, Zhang Z, Wang W, Lu H, Wang Y, Liao J (2022) Timereplayer: Unlocking the potential of event cameras for video interpolation. In: CVPR, pp 17804–17813

IniVation (2020) Understanding the performance of neuromorphic event-based vision sensors. https://inivationcom/

Jiang H, Sun D, Jampani V, Yang MH, Learned-Miller E, Kautz J (2018) Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In: CVPR, pp 9000–9008

Jin M, Hu Z, Favaro P (2019) Learning to extract flawless slow motion from blurry videos. In: CVPR, pp 8112–8121

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv:14126980

Lao Y, Ait-Aider O (2018) A robust method for strong rolling shutter effects correction using lines with automatic feature selection. In: CVPR, pp 4795–4803

Li M, Wang P, Zhao L, Liao B, Liu P (2024) Usb-nerf: Unrolling shutter bundle adjusted neural radiance fields. In: ICLR

Lichtsteiner P, Posch C, Delbruck T (2008) A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. IEEE J Solid-State Circuits 43(2):566–576, DOI 10.1109/JSSC.2007.914337

Lin S, Zhang J, Pan J, Jiang Z, Zou D, Wang Y, Chen J, Ren J (2020) Learning event-driven video deblurring and interpolation. In: ECCV, pp 695–710

Lin X, Li Y, Zhu J, Zeng H (2023) Deflickercyclegan: Learning to detect and remove flickers in a single image. IEEE TIP 32:709–720

Liu P, Cui Z, Larsson V, Pollefeys M (2020) Deep shutter unrolling network. In: CVPR, pp 5941–5949

Liu S, Yuan L, Tan P, Sun J (2013) Bundled camera paths for video stabilization. ACM TOG 32(4):1–10

Liu Z, Yeh RA, Tang X, Liu Y, Agarwala A (2017) Video frame synthesis using deep voxel flow. In: ICCV, pp 4463–4471

Loshchilov I, Hutter F (2016) Sgdr: Stochastic gradient descent with warm restarts. arXiv:160803983

Lu Y, Liang G, Wang L (2023) Self-supervised learning of event-guided video frame interpolation for rolling shutter frames. arXiv:230615507

Niklaus S, Mai L, Liu F (2017a) Video frame interpolation via adaptive convolution. In: CVPR, pp 670–679

Niklaus S, Mai L, Liu F (2017b) Video frame interpolation via adaptive separable convolution. In: ICCV, pp 261–270

Niu M, Chen T, Zhan Y, Li Z, Ji X, Zheng Y (2025) Rs-nerf: Neural radiance fields from rolling shutter images. In: ECCV, pp 163–180

Oth L, Furgale P, Kneip L, Siegwart R (2013) Rolling shutter camera calibration. In: CVPR, pp 1360–1367

Pan L, Scheerlinck C, Yu X, Hartley R, Liu M, Dai Y (2019) Bringing a blurry frame alive at high frame-rate with an event camera. In: CVPR, pp 6820–6829

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al (2019) Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS

Purkait P, Zach C (2018) Minimal solvers for monocular rolling shutter compensation under ackermann motion. In: WACV, pp 903–911

Purkait P, Zach C, Leonardis A (2017) Rolling shutter correction in manhattan world. In: ICCV, pp 882–890

Rebecq H, Gehrig D, Scaramuzza D (2018) Esim: an open event camera simulator. In: Conference on Robot Learning, pp 969–982

Reda FA, Sun D, Dundar A, Shoeybi M, Liu G, Shih KJ, Tao A, Kautz J, Catanzaro B (2019) Unsupervised video interpolation using cycle consistency. In: ICCV, pp 892–900

Rengarajan V, Rajagopalan AN, Aravind R (2016) From bows to arrows: Rolling shutter rectification of urban scenes. In: CVPR, pp 2773–2781

Rengarajan V, Balaji Y, Rajagopalan A (2017) Unrolling the shutter: Cnn to correct motion distortions. In: CVPR, pp 2291–2299

Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp 234–241

Schonberger JL, Frahm JM (2016) Structure-from-motion revisited. In: CVPR, pp 4104–4113

Sheinin M, Chan D, O'Toole M, Narasimhan SG (2022) Dual-shutter optical vibration sensing. In: CVPR, pp 16324–16333

Sun D, Yang X, Liu MY, Kautz J (2018) Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: CVPR, pp 8934–8943

Tulyakov S, Gehrig D, Georgoulis S, Erbach J, Gehrig M, Li Y, Scaramuzza D (2021) Time lens: Event-based video frame interpolation. In: CVPR, pp 16155–16164

Tulyakov S, Bochicchio A, Gehrig D, Georgoulis S, Li Y, Scaramuzza D (2022) Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In: CVPR, pp 17755–17764

Wang B, He J, Yu L, Xia GS, Yang W (2020) Event enhanced high-quality image recovery. In: ECCV, pp 155–171

Xu F, Yu L, Wang B, Yang W, Xia GS, Jia X, Qiao Z, Liu J (2021) Motion deblurring with real events. In: ICCV, pp 2583–2592

Zhang D, Ding Q, Duan P, Zhou C, Shi B (2022) Data association between event streams and intensity frames under diverse baselines. In: ECCV, pp 72–90

Zhang X, Yu L (2022) Unifying motion deblurring and frame interpolation with events. In: CVPR, pp 17765–17774

Zhang X, Huang H, Jia X, Wang D, Zhang L, Zheng B, Zhou W, Lu H (2024) Neural image re-exposure. Computer Vision and Image Understanding 248:104094

Zhong Z, Zheng Y, Sato I (2021) Towards rolling shutter correction and deblurring in dynamic scenes. In: CVPR, pp 9219–9228

Zhong Z, Cao M, Sun X, Wu Z, Zhou Z, Zheng Y, Lin S, Sato I (2022) Bringing rolling shutter images alive with dual reversed distortion. In: ECCV, pp 233–249

Zhou X, Duan P, Ma Y, Shi B (2022) Evunroll: Neuromorphic events based rolling shutter image correction. In: CVPR, pp 17775–17784

Zhu AZ, Yuan L, Chaney K, Daniilidis K (2018) Ev-flownet: Self-supervised optical flow estimation for event-based cameras. arXiv:180206898

Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV, pp 2223–2232

Zhuang B, Cheong LF, Hee Lee G (2017) Rolling-shutter-aware differential sfm and image rectification. In: ICCV, pp 948–956

Zhuang B, Tran QH, Ji P, Cheong LF, Chandraker M (2019) Learning structure-and-motion-aware rolling shutter correction. In: CVPR, pp 4551–4560