A Unified Framework for Event-based Frame Interpolation with Ad-hoc Deblurring in the Wild

Lei Sun, Daniel Gehrig, Christos Sakaridis, Mathias Gehrig, Jingyun Liang, Peng Sun, Zhijie Xu, Kaiwei Wang, Luc Van Gool, and Davide Scaramuzza

Abstract—Effective video frame interpolation hinges on the adept handling of motion in the input scene. Prior work acknowledges asynchronous event information for this, but often overlooks whether motion induces blur in the video, limiting its scope to sharp frame interpolation. We instead propose a unified framework for event-based frame interpolation that performs deblurring ad-hoc and thus works both on sharp and blurry input videos. Our model consists in a bidirectional recurrent network that incorporates the temporal dimension of interpolation and fuses information from the input frames and the events adaptively based on their temporal proximity. To enhance the generalization from synthetic data to real event cameras, we integrate self-supervised framework with the proposed model to enhance the generalization on real-world datasets in the wild. At the dataset level, we introduce a novel real-world high-resolution dataset with events and color videos named HighREV, which provides a challenging evaluation setting for the examined task. Extensive experiments show that our network consistently outperforms previous state-of-the-art methods on frame interpolation, single image deblurring, and the joint task of both. Experiments on domain transfer reveal that self-supervised training effectively mitigates the performance degradation observed when transitioning from synthetic data to real-world data.

Index Terms—Event camera, video frame interpolation, motion deblurring, self-supervised learning, low-level vision

1 Introduction

Video *testtest* frame interpolation (VFI) methods synthesize intermediate frames between consecutive input frames, increasing the frame rate of the input video, with wide applications in super-slow generation [19], [22], [32], video editing [40], [62], virtual reality [1], and video compression [56]. With the absence of inter-frame information, frame-based methods explicitly or implicitly utilize motion models such as linear motion [22] or quadratic motion [57]. However, the non-linearity of motion in real-world videos makes it hard to accurately capture inter-frame motion with these simple models.

Recent works introduce event cameras in VFI as a proxy to estimate the inter-frame motion between consecutive frames. Event cameras [14] are bio-inspired asynchronous sensors that report per-pixel intensity changes, *i.e.*, *events*, instead of synchronous full intensity images. The events are recorded at high temporal resolution (in the order of μs) and high dynamic range (over 140 dB) within and between frames, providing valid compressed motion information. Previous works [17], [52], [53] show the potential of event cameras in VFI, comparing favorably to frame-only methods, especially in high-speed non-linear motion scenarios, by using spatially aligned events and RGB frames. These event-based VFI methods make the crucial assumption that the input images are sharp. However, this assumption is violated in real-world scenes because of the ubiquitous motion blur. In particular, because of the finite exposure time of frames in real-world videos, especially of those captured with event cameras that output both image frames and an event stream (*i.e.*, Dynamic and Activate VIsion Sensor (DAVIS) [5])—which have a rather long exposure time and low frame rate, motion blur is inevitable for high-speed scenes. In such a scenario, where the reference frames for VFI are degraded by motion blur, the performance of frame interpolation also degrades.

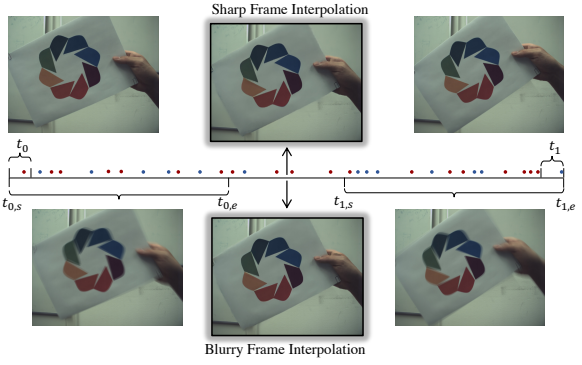


Fig. 1. Our unified framework for event-based sharp and blurry frame interpolation. Red/blue dots: negative/positive events; Curly braces: exposure time range.

As events encode motion information within and between frames, several studies [9], [29], [35] are carried out on event-based deblurring in conjunction with VFI. However, these works

- *L. Sun, P. Sun, and K. Wang are with the National Research Center for Optical Instrumentation, Zhejiang University, 310027 Hangzhou, China. E-mails: {leo_sun, pengsunr, wangkaiwei}@zju.edu.cn.*
- *L. Sun, D. Gehrig, M. Gehrig, and D. Scaramuzza are with the Robotics and Perception Group, University of Zurich, 8050 Zurich, Switzerland. E-mails: leo_sun@zju.edu.cn, {dgehrig, mgehrig}@ifi.uzh.ch.*
- *L. Sun, C. Sakaridis, J. Liang, and L. Van Gool are with the Computer Vision Lab, ETH Zürich, 8092 Zürich, Switzerland. E-mails: {leisun, csakaridis, jinliang, vangool}@vision.ee.ethz.ch.*
- *Zhijie Xu is with the Centre for Visual and Immersive Computing, Huddersfield University, HD1 3DH UK. E-mail: z.xu@hud.ac.uk.*

- *This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No.12174341, in part by the National Key R&D Program of China under Grant No.2022YFF0705500 and No.2022YFB3206000, in part by the China Scholarship Council, in part by AlpsenTek GmbH.*

- *Corresponding author: K. Wang.*

approach the problem via cascaded deblurring and interpolation pipelines and the performance of VFI is limited by the image deblurring performance.

Thus, the desideratum in event-based VFI is robust performance on both sharp image interpolation and blurry image interpolation. Frame-based methods [21], [26], [26], [33], [44], [63] usually treat these two aspects as separate tasks. Different from frames, events are not subject to motion blur. No matter whether the frame is sharp or blurry, the corresponding events are the same. Based on this observation, we propose to unify the two aforementioned tasks into one problem: *given two input images and a corresponding event stream, restore the latent sharp images at arbitrary times between the input images.* The input images could be either blurry or sharp, as Fig. 1 shows. To solve this problem, we first revisit the physical model of event-based deblurring and frame interpolation. Based on this model, we propose a novel recurrent network, named **R**ecurrent **E**vent-based **F**rame **I**nterpolation with ad-hoc **D**eblurring (REFID), which can perform event-based sharp VFI, event-based blurry VFI, and single image deblurring. The network consists of two branches, an image branch and an event branch. The recurrent structure pertains to the event branch, in order to enable the propagation of information from events across time in both directions. Features from the image branch are fused into the recurrent event branch at multiple levels using a novel attention-based module for event-image fusion, which is based on the squeeze-and-excitation operation [18].

In the realm of event-based image and video deblurring (*i.e. single image deblurring* and *blurry frame interpolation* in our work), a significant challenge lies in achieving robust generalization to real-world conditions [20], [37]. The distinctive characteristics arising from hardware limitations, sensor noise, and uncertainties in parameters such as the threshold (c in (1)) contribute to a domain gap between synthetic events and those captured by specific event cameras. While some researchers address this issue by collecting real-world data [50], the impracticality of obtaining ground truth for each event camera poses a substantial limitation. To tackle this challenge, we propose a novel approach involving the fine-tuning of the model on real-world data, leveraging a self-supervised learning methodology. This approach facilitates generalization across different cameras without the need for collecting ground truth. Within the framework of our proposed self-supervised learning approach, we integrate various constraints throughout the image degradation process and delve into the exploration of motion compensation in the domain of event-based deblurring.

To test our method on a real-world setting and motivated by the lack of event-based datasets recorded with high-quality event cameras, we record a dataset, HighREV, with high-resolution chromatic image sequences and corresponding events. From the sharp image sequences, we synthesize blurry images by averaging several consecutive frames [31]. To our knowledge, HighREV has the highest resolution in both image and event among all publicly available event datasets.

In summary, we make the following contributions:

- We propose a framework for solving general event-based frame interpolation and event-based single image deblurring, which builds on the underlying physical model of high-frame-rate video frame formation and event generation.
- We introduce a novel network for solving the above tasks, which is based on a bi-directional recurrent architecture, includes an event-guided channel-level attention fusion module that adaptively attends to features from the two input frames according to the temporal proximity with features from the event branch, and achieves state-of-the-art results on both synthetic and real-world datasets.
- We integrate REFID with a self-supervised framework with motion compensation for event-based image/video deblurring. This utilizes the event generative model and constraints between the blurry images and a sharp video clip.
- We evaluate the proposed framework on benchmarks with synthetic events and real events in self-supervised learning settings. The integration of REFID with a self-supervised fine-tuning framework allows for model refinement using real data, even in the absence of ground truth.
- We present a new real-world high-resolution dataset with events and RGB videos, which enables real-world evaluation of event-based interpolation and deblurring.

2 Related Work

2.1 Event-based frame interpolation

Because event cameras report the per-pixel intensity changes, they provide useful spatio-temporal information for frame interpolation. Tulyakov *et al*. [53] propose Time Lens, which combines a warping-based method and a synthesis-based method with a late-fusion module. Time Lens++ [52] further improves the efficiency and performance via computing motion splines and multi-scale fusion separately. TimeReplayer [17] utilizes a cycle-consistency loss as supervision signal, making a model trained on low-frame-rate videos also able to predict high-speed videos. All the methods above assume that the key frame is sharp, but in high-speed or low-illumination scenarios, the key frame inevitably gets blurred because of the high-speed motion within the exposure time, where these methods failed (Tab. 2). Hence, the exposure time should be taken into consideration in real-world scenes.

2.2 Event-based deblurring

Due to the high temporal resolution, event cameras provide motion information within the exposure time, which is a natural motion cue for image deblurring. Thus, several works have focused on event-based image deblurring. Jiang *et al*. [23] used convolutional models and mined the motion information and edge information to assist deblurring. Sun *et al*. [50] proposed a multi-head attention mechanism for fusing information from both modalities, and designed an event representation specifically for the event-based image deblurring task. Kim *et al*. [24] further extended the task to images with unknown exposure time by activating the events that are most related to the blurry image. These methods only explore the single image deblurring setting, where the timestamp of the deblurred image is in the middle of the exposure time. However, the events encode motion information for the entire exposure time, and latent sharp images at arbitrary points within the exposure time can be estimated in theory.

2.3 Joint frame interpolation and enhancement

Pan *et al*. [35] formulate the Event Double Integral (EDI) deblurring model, which is derived from the definition of image blur and the measurement mechanism of event cameras, and perform both image deblurring and frame interpolation by accumulating events and applying the intensity changes within the exposure time and from the key frame to the synthesized frames, respectively.

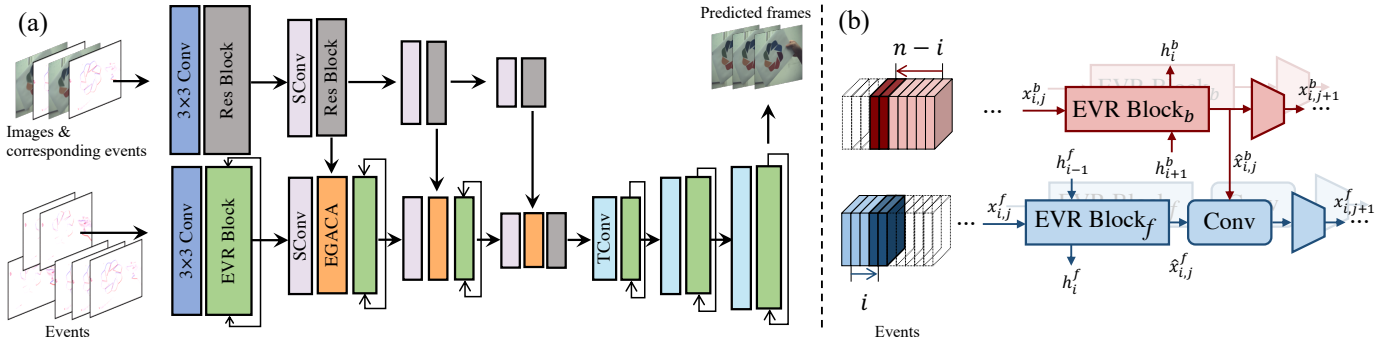


Fig. 2. (a): The architecture of our Recurrent Event-based Frame Interpolation with ad-hoc Deblurring (REFID) network. The input of the image branch consists of two key frames and their corresponding events, and the event branch consumes sub-voxels of events recurrently. "EGACA": event-guided adaptive channel attention, "SConv": strided convolution, "TConv": transposed convolution. (b): The proposed bidirectional event recurrent (EVR) blocks. In each recurrent step, the events from the forward and backward direction are fed to the network. For notations, cf. (10).

This seminal work optimizes the model by minimizing an energy function but is limited by practical issues in the measurement mechanism of event cameras, e.g. accumulated noise, dynamic contrast thresholds and missing events. Based on EDI, a differentiable model and a residual learning denoising model to improve the result is introduced in [55]. Recent works [3], [36], [61] identify the relationship between the events and the latent sharp image, and apply it to self-supervised event-based image reconstruction and image deblurring. However, the above works on joint frame interpolation and deblurring predict the latent frames with a two-stage deblurring+interpolation approach, which limits the performance of VFI.

2.4 Self-supervised deblurring

Traditional deblurring methods model the image degradation process as a convolution of a blur kernel with a latent sharp image, in which estimating accurate blur kernels is essential to the result. Other researchers dedicated lots of efforts to designing regularizations to deal with the highly ill-posed problem: total-variance [4], Gaussian distribution [27], intensity and gradient prior of text images [34], *etc*. However, these assumptions may fail in complex real-world scenarios. Other works [7], [12], [13], [28], [42] attempted to remove the dependence on assumptions by utilizing spatially variant blur kernels. However, these methods fail to account for non-planar and object-moving scenes, making them unsuitable for practical use. Chen *et al*. [8] predicted the optical flow using latent images from kernel-free estimation and re-render the blurry image, and combining self-supervised losses with ground truth for a hybrid training to improve the deblurring results. Liu *et al*. [30] constrained the self-supervised image deblurring with the linear motion assumption in the exposure time and proposed a differentiable model to complete the blur consistency with predicted optical flow and latent sharp images. Based on [30], Yu *et al*. [58] proposed to utilize events to predict the optical flow within the exposure time of the image, and performed single image deblurring. In contrast to Yu et al. [58], who solely considered events within the exposure time, our methods incorporate an event-generation model and leverage event information both within and beyond the exposure time.

3 Our Approach

We first revisit the physical model of event-based frame interpolation and deblurring in Sec. 3.1. Based on this model, we argue that the events within the exposure time should not be ignored in event-based frame interpolation, and present our model architecture abstracted from the physical model in Sec. 3.2. To perform the bidirectional recurrent propagation, we demonstrate the data preparation in Sec. 3.3. In Sec. 3.4 and Sec 3.5, we introduce the proposed bidirectional Event Recurrent Block and Event-Guided Adaptive Channel Attention in detail. Note that all the symbols are summarized in Tab. 8 in the Supplementary Materials.

3.1 Problem Formulation

Once the change in intensity $\mathcal{I}$ at a pixel between the current moment and the moment of the last generated event at that pixel surpasses the contrast threshold c, an event camera emits the i-th event e_i, represented as a tuple $e_i = (x_i, y_i, t_i, p_i)$, where x_i and y_i represent the pixel coordinates of the event, t_i represents its timestamp, and p_i is the polarity of the single event. More formally, this can be written as

$$p_i = \begin{cases} +1, \text{if} \log\left(\frac{\mathcal{I}_t(x_i,y_i)}{\mathcal{I}_{t-\Delta t}(x_i,y_i)}\right) > c, \\ -1, \text{if} \log\left(\frac{\mathcal{I}_t(x_i,y_i)}{\mathcal{I}_{t-\Delta t}(x_i,y_i)}\right) < -c. \end{cases} \tag{1}$$

Ideally, given two consecutive images, referred to as the left frame I_0 and the right frame I_1, and the corresponding event stream in the time range between the timestamps of the two images $[t_0, t_1]$, we can get any latent image $\hat{I}_\tau$ with timestamp τ in $[t_0, t_1]$ via

$$\begin{aligned} \hat{I}_\tau &= I_0 \exp(c \int_{t_0}^{t} p(s)ds), \\ \hat{I}_\tau &= I_1 \exp(c \int_{t_1}^{t} p(s)ds), \end{aligned} \tag{2}$$

where $p(s)$ is the polarity component of the event stream. Note that in (2), the $p(s)$ is the set of all the polarities of the event stream, with the same dimension as the event image.

Previous event-based methods [17], [52], [53] solve event-based frame interpolation based on (2). However, in the real-world setting, because of the finite exposure times of the two frames, the timestamps t_0 and t_1 should be replaced by time ranges, and the images I_0 and I_1 may be either sharp (small motion in the exposure time) or blurry (large motion in the exposure time). Thus, the events within the exposure time of the frames, e, should also be utilized for removing potential blur from the frames:

$$\text{Deblur}(I, e) = \frac{B \times T}{\int_{t_s}^{t_e} \exp\left(c \int_{\frac{t_s+t_e}{2}}^{t} p(s)ds\right) dt}, \tag{3}$$

where B, T, t_s and t_e are the blurry frame, length of exposure time, start and end of exposure time, respectively. Previous studies [9], [23], [29], [35] combine the above deblur equation with the frame interpolation equation (2) (denoted as Interpo) to synthesize the target frame:

$$\begin{aligned} \hat{I}_{\tau,0} &= \text{Deblur}(I_0, \epsilon_{t_{0,s}\to t_{0,e}})\text{Interpo}(\epsilon_{t_{0,s}\to\tau}), \\ \hat{I}_{\tau,1} &= \text{Deblur}(I_1, \epsilon_{t_{1,s}\to t_{1,e}})\text{Interpo}(\epsilon_{\tau\leftarrow t_{1,e}}), \end{aligned} \tag{4}$$

where $\epsilon_{t_{0,s}\to\tau}$ and $\epsilon_{\tau\leftarrow t_{1,e}}$ indicate the intensity changes—recorded as events—from the start of the exposure time of the left frame, $t_{0,s}$, and the end of the exposure time of the right frame, $t_{1,e}$, to the target timestamp τ.

However, the physical model (4) is prone to sensor noise and to the varying contrast threshold of an event camera, which is an inherent drawback of such cameras.

Based on (4), [9], [29], [35], [61] design deep neural networks with a cascaded *first-deblur-then-interpolate* pipeline to perform blurry frame interpolation. In these two-stage methods, the performance of frame interpolation (second stage) is limited by the performance of image deblurring (first stage). Moreover, these methods are only evaluated on blurry frame interpolation.

Given the left and right frame, we design a unified framework to perform event-based frame interpolation both for sharp and blurry inputs with a one-stage model, which applies deblurring ad-hoc.

3.2 General Architecture

The physical model of (4) indicates that the latent sharp frame at time τ can be derived from the two consecutive frames and the corresponding events as

$$\begin{aligned} \hat{I}_{\tau,0} &= \mathbf{F}(\mathbf{G}(I_0, E_0), E_{t_{0,s}\to\tau}), \\ \hat{I}_{\tau,1} &= \mathbf{F}(\mathbf{G}(I_1, E_1), E_{\tau\leftarrow t_{1,e}}), \end{aligned} \tag{5}$$

where $\mathbf{G}$ and $\mathbf{F}$ are learned parametric mappings, representing "Deblur" and "Interpo" function in (4). Contrary to the formulation of (4), $\mathbf{G}$ does not accomplish solely image deblurring, but rather extracts features of both absolute intensities (image) and relative intensity changes (events) within the exposure time. We use cascaded residual blocks to model this mapping, *i.e.*, "Res Block" in Fig. 2. For each latent frame, previous methods collect the events in both time ranges and convert them to an event representation [17], [53], which may incur inconsistencies in the result [52]. To mitigate this, we use a recurrent network to naturally model temporal information. Thus, we abstract the physical model (4) to:

$$\begin{aligned} \hat{I}_{\tau,0} &= \mathbf{EVR}_f(\mathbf{G}(I_0, E_0, I_1, E_1), E_\tau, E_{t_{0,s}\to\tau}), \\ \hat{I}_{\tau,1} &= \mathbf{EVR}_b(\mathbf{G}(I_0, E_0, I_1, E_1), E_\tau, E_{\tau\leftarrow t_{1,e}}), \end{aligned} \tag{6}$$

where $\mathbf{EVR}_f$ and $\mathbf{EVR}_b$ denote forward and backward event recurrent (EVR) blocks, respectively. (6) summarizes the architecture of our proposed method. E_τ refers to the events in a small time range centered around τ. The recurrent blocks accept as input not only current events, but also previous event information through their hidden states.

Because of the sensor noise and the varying contrast threshold of the event camera sensor, $\hat{I}_{\tau,0}$ ($\hat{I}_{\tau,1}$) approximates the latent sharp image more accurately when the corresponding timestamp of the latter, τ, is closer to t_0 (t_1). To fuse $\hat{I}_{\tau,0}$ and $\hat{I}_{\tau,1}$ implicitly, we further propose a new Event-Guided Adaptive Channel Attention (EGACA) module to mine and fuse the features from the image branch of REFID with adaptive weights determined by the current events:

$$\hat{I}_\tau = \text{Fuse}(\hat{I}_{\tau,0}, \hat{I}_{\tau,1}). \tag{7}$$

The overall network architecture of REFID is shown in Fig. 2 (a). The image branch extracts features from the two input images and the corresponding events and is connected to the event branch at multiple scales. Overall, REFID has a U-Net [41] structure. A bidirectional recurrent encoder with EVR blocks extracts features from current events and models temporal relationships with previous and future events. In each block of the encoder, the features from the image branch are fused with those from the event branch adaptively with our novel EGACA module, which we detail in 3.5.

The proposed REFID can be extended to single-image deblurring by utilizing a single frame and its corresponding events for the image and event branches, respectively. Moreover, our approach generates multiple latent sharp images as opposed to only one, as highlighted in the work by Sun et al. [50].

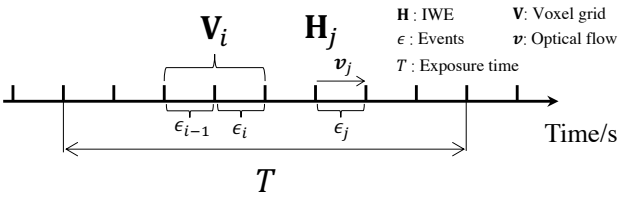


Fig. 3. Details of network inputs. Events within the exposure time T and the blurry frame are unfolded into N sharp images. Events are split into sub-intervals ϵ_i, and two sub-intervals of events are used to compute 2-channel voxel grids $\mathbf{V}_i$. ϵ_i is also used to predict optical flow $\mathbf{u}_j$. Events are warped to produce IWE $\mathbf{H}_j$ with $\mathbf{u}_j$ for each sub-interval.

3.3 Data Preparation

To feed the asynchronous events to our network, we first need to convert them to a proper representation. The detailed model inputs are depicted in Fig. 3. According to (6), the latent image can be derived in both temporal directions. Thus, apart from the forward event stream, we reverse the event stream both in time and polarity to get a backward event stream. Then, event streams from the two directions are converted to two voxel grids [38], [39], and we take one voxel grid $\mathbf{V}_{total} \in \mathbb{R}^{(n+2)\times H\times W}$ for an example, where n is the number of interpolated frames $V_i \in \mathbb{R}^{(n+2)\times H\times W}$ and $V \in \mathbb{R}^{(n+2)\times H\times W}$, where n is the number of interpolated frames. The channel dimension of the voxel grids holds discrete temporal information. In each recurrent iteration, $\mathbf{V}_\text{i} \in \mathbb{R}^{2\times H\times W}$ are constructed from small sub-intervals of events $\epsilon_i = \{e_k | \tau_i \le t_k \le \tau_{i+1}\}$. $\mathbf{V}_\text{i}$ from both directions are fed to the event branch, which encodes the event information for the latent frame. We also convert events in the exposure time of the two images to voxel grids and concatenate them with corresponding images to form the input of the image branch.

Further, for self-supervised fine-tuning experiments, each group of events is also used to compute an optical flow segment in self-supervised settings $\mathbf{v}(\mathbf{x}_k)$ following the model-based method [47] for event-based optical flow estimation and a 2-channel image of warped events (IWE) $\mathbf{H}_i(\mathbf{x})$ with the estimated

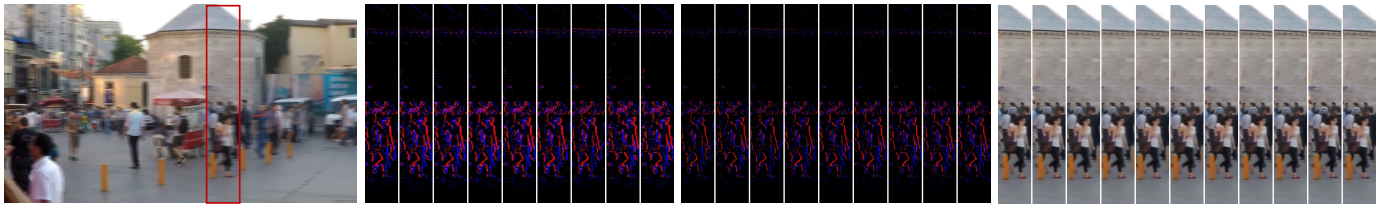


Fig. 4. A example for self-supervised single-image deblurring, a sharp video clip is restored with a blurry image and corresponding events. From left to right: Visualized events, image of warped events (IWE), and resulting sharp video clip. IWE provides sharper edge information while events contribute to capturing the blurry shape information.

optical flow, where $\mathbf{x} \doteq (x, y)^\top$. For IWE we compute the projection of motion-compensated events via

$$\mathbf{H}_i(\mathbf{x}) = \sum_{e_k \in \epsilon_i} \delta(p - p_k)\delta(\mathbf{x} - \mathbf{u}'_k), \tag{8}$$

$$\mathbf{u}'_k = \mathbf{u}_k - \mathbf{v}_i(\mathbf{u}_k)(t_k - \tau_i), \tag{9}$$

where δ represents the Kronecker delta. With the equation we temporally align all events with the interval to the timestamp τ_i[1]. Note that the result is a 2-channel tensor, where events are separated according to their polarity. As Fig. 4 shows, the image of warped events provides a strong inductive bias for our network by showing it where sharp edges are to be expected. To compute the 2-channel voxel grid input $\mathbf{V}_i(\mathbf{x})$, we concatenate two 1-channel voxel grids computed from events ϵ_{i-1} and ϵ_i[2]. Before passing to the network, we concatenate the voxel grid and image of warped events, resulting in a 4-channel input, and apply input normalization.

3.4 Bidirectional Event Recurrent Block

In previous event-based works [17], [52], [53], for each latent sharp image, the events from both left and right images to the target image are accumulated and converted to an event representation. However, compared to the temporal resolution of events, the length of the exposure time of frames is large and not negligible, so simple accumulation from a single timestamp in the above works loses information and is not reasonable. Moreover, inference for different latent frames is segregated, which leads to inconsistencies in the results [52]. To deal with these problems, we propose a recurrent architecture that models the temporal information both within the exposure time of each frame and between exposure times of different frames. By adopting recurrent blocks, frame interpolation is independent from the exposure time of key frames and it can also be performed inside the exposure time. Features propagated through hidden states of the network also guarantee consistency across the predicted frames. Based on (5), we design a bidirectional Event Recurrent (EVR) block to iteratively extract features from the event branch. As Fig. 2 (b) shows, for each direction, the input sub-voxel $(i-1, i)$ only consists of two voxels of the input voxel. In the next recurrent iteration, the selected sub-voxel moves forward to the next time $(i, i+1)$. For a given recurrent iteration i, the forward EVR block cycles for i times and the backward EVR block cycles for $n-i$ times, where n is the index of the latent sharp image at hand:

$$\begin{aligned}
\hat{x}^b_{i,j+1}, h^b_i &= \mathbf{EVR}_b(x^b_{i,j}, h^b_{i+1}),\\
\hat{x}^f_{i,j+1}, h^f_i &= \mathbf{EVR}_f(x^f_{i,j}, h^f_{i-1}),\\
x^b_{i,j+1} &= \mathrm{Down}(\hat{x}^b_{i,j+1}),\\
x^f_{i,j+1} &= \mathrm{Down}(\mathrm{Conv}(\mathrm{Concat}(\hat{x}^b_{i,j+1}, \hat{x}^f_{i,j+1}))),
\end{aligned} \tag{10}$$

where i and j are the indices of sub-voxel and scale, respectively. x, h, f and b denote feature flow, hidden state, forward and backward, respectively. We select ResNet as the architecture for the EVR block instead of ConvLSTM [45] or ConvGRU [46], because the time range of events between consecutive frames is rather short (cf. Tab 5). In each EVR block, the features from the two directions are fused through convolution and downsampled to half of the original size. The bidirectional EVR blocks introduce the information flow from both directions, which models $E_{t_{0,s} \to \tau}$ and $E_{\tau \leftarrow t_{1,e}}$ in (5), helping reduce artifacts by using the information from the end of the time interval (cf. Fig. 9).

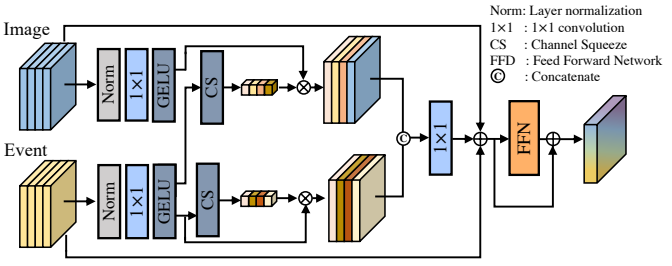


Fig. 5. The Event-Guided Adaptive Channel Attention module. The channel weights for the image branch are extracted from the event branch.

3.5 Event-Guided Adaptive Channel Attention

In event-based frame interpolation, fusion happens both between the two input frames and between frames and events. Because of the inherent noise of event cameras, the longer the time range between the key frames is, the more the noise in the event accumulation increases. Ideally, the key frame that is closer to the latent frame should contribute more to the prediction of the latter. In other words, the weights of two key frames should be decided by *time*.

In our REFID network, the two key frames and the corresponding events are concatenated along the channel dimension to provide the input of the image branch. We design the novel Event-Guided Adaptive Channel Attention (EGACA) module to

1. Note, that for the last interval ϵ_{N-1}, we generate two images of warped events $\mathbf{H}_{N-1}(\mathbf{x})$ and $\mathbf{H}_N(\mathbf{x})$, by once backward warping, and then forward warping the same events.

2. Note that this includes two channels that go beyond the exposure interval, ϵ_0 and ϵ_N

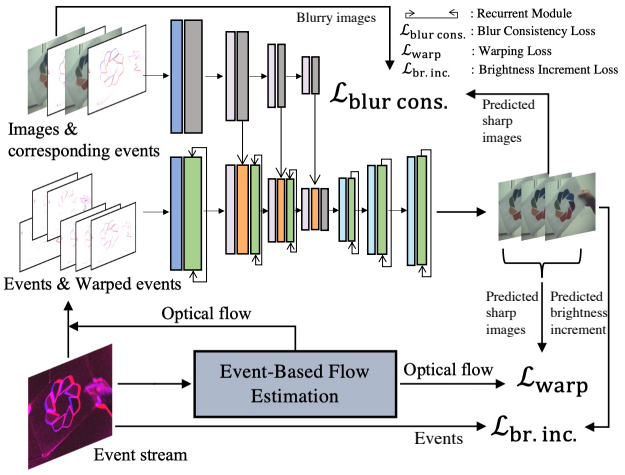


Fig. 6. Self-supervised framework with the same basic architecture illustrated in Fig 2: The events are directed to the optical flow estimation module, with the resulting flow utilized for warping events for IWE and computing the Warping Loss with (18). Input blurry images and predicted sharp images are utilized in the calculation of Blur Consistency Loss in (17). Input events and predicted brightness increment are used in the calculation of Brightness Increment Loss in (13).

fuse the two key frames and events at the current input sub-voxel in the recurrent structure. The current input sub-voxel contains events in a small range around the timestamp of the latent frame and the fusion weights for the two key frames and the events are determined by the current input sub-voxel, which indicates the time.

Fig. 5 shows the detailed architecture of the proposed EGACA. We simplify the multi-head channel attention of EFNet [50] to channel attention from SENet [18]. Two Channel Squeeze (CS) blocks extract channel weights from the current events, and two weights multiply event features and image features for self-attention and event-guided attention to image features, respectively. Then, feature maps from the two branches are fused by a feed-forward network. In each recurrent iteration, the channel weights from the current events are different, which helps to mine different features from the two images along the channel dimension.

3.6 Self-supervised Framework

Fig. 6 shows the framework of the self-supervised fine-tuning strategy. In the self-supervised fine-tuning, we find that three self-supervised loss functions are sufficient to train our model, namely (*i*) a brightness increment loss, (*ii*) a blur consistency loss, and (*iii*) a warping loss. While the blur consistency loss is calculated over the entire interval, the other two losses can be defined for each time interval at τ_i and thus we obtain dense supervision throughout the interval. The resulting loss contains the following $1 + 2(N-1)$ terms

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{\text{blur cons.}} + \sum_{i=1}^{N-1} \lambda_0 \mathcal{L}^i_{\text{br. inc.}} + \lambda_1 \mathcal{L}^i_{\text{warp}}, \tag{11}$$

and we will now discuss these in turn. Note that all the self-supervised experiments are conducted with single-image deblurring settings

Brightness Increment Loss As alluded to in (1), event cameras approximately measure (up to quantization) the brightness increment between two time instances. We enforce this constraint by minimizing the difference between the measured and predicted log brightness increment between adjacent predicted frames: $\Delta\hat{L}_i(\mathbf{x}) \doteq \log \hat{I}_{i+1}(x) - \log \hat{I}_i(x)$.

$$\mathcal{L}_{\text{br. inc.}} = \|\Delta L(\mathbf{x}; \epsilon_i) - \Delta\hat{L}_i(\mathbf{x})\|_2^2. \tag{12}$$

We compute the brightness increment as in (1). Since ΔL depends on an unknown contrast threshold c, we slightly modify the above loss, to minimize the difference of normalized terms:

$$\mathcal{L}^i_{\text{br. inc.}} = \left| \frac{\Delta L(\mathbf{x}; \epsilon_i)}{\|\Delta L(\mathbf{x}; \epsilon_i)\|_2} - \frac{\Delta\hat{L}_i(\mathbf{x})}{\|\Delta\hat{L}_i(\mathbf{x})\|_2} \right|. \tag{13}$$

This modification results in a cancellation of the constant term c. Note that this is a common trick employed in works like [6], [15], [36]. However, different from these works, we do not minimize the first-order linear approximation to the event generation model, but instead use the true model without linearization, and thus are free from approximation error.

Blur Consistency Loss: Theoretically, the intensity of the blurry image equals the average intensity of all the latent sharp images within the exposure time of the blurry image:

$$B(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} I_i(\mathbf{x}). \tag{14}$$

Based on the image blurring process above, we devise a self-supervised loss that relates the measured blurry frame and predicted sharp frames via

$$\mathcal{L}_{\text{blur cons.}} = \left\| B(\mathbf{x}) - \frac{1}{N} \sum_{i=1}^{N} \hat{I}_i(\mathbf{x}) \right\|_2^2. \tag{15}$$

Here all the I are within the exposure time of the blurry image B. We found that $N = 11$ does not generate sufficiently many sharp images to generate realistic blur, and for this reason, we reuse the optical flow derived from [47] (a non-learning method) directly to upsample the frames by a factor of $M = 4$, generating 3 additional intermediate frames between consecutive sharp images (Note that the event-based flow estimation module is not the contribution of our work, and it can any other event-based optical flow estimation method):

$$\hat{I}_{i,m}(\mathbf{x}) = \hat{I}_{i+1}\left(\mathbf{x} - \frac{m}{M} v_i(\mathbf{x})\right), \quad \text{for } m = 1, ..., M-1. \tag{16}$$

For each frame at time τ_i we get intensities from the next frame at τ_{i+1} by rescaling the flow appropriately and applying bilinear sampling. as a result, the blur consistency loss becomes

$$\mathcal{L}_{\text{blur cons.}} = \left\| B(\mathbf{x}) - \frac{1}{NM} \sum_{i=1}^{N} \sum_{m=0}^{M-1} \hat{I}_{i,m}(\mathbf{x}) \right\|_2^2, \tag{17}$$

where we make use of the fact that $\hat{I}_{i,0} = \hat{I}_{i+1}$.

Warping Loss: With the dense optical flow $\mathbf{v}_i(\mathbf{x})$ recovered previously, we warp the latent sharp image back to the last timestamp and compare the warped image and predicted image. With the dense optical flow, this loss is able to also constrain slight intensity changes.

$$\mathcal{L}^i_{\text{warp}} = \left\| \hat{I}_i(\mathbf{x}) - \hat{I}_{i+1}(\mathbf{x} - \mathbf{v}_i(\mathbf{x})) \right\|_2^2. \tag{18}$$

Note that for the blurry frame interpolation task, the blur consistency loss is applied to the target frame with the timestamp in the exposure time of either blurry frame. For the single image deblurring, all three losses are applied.

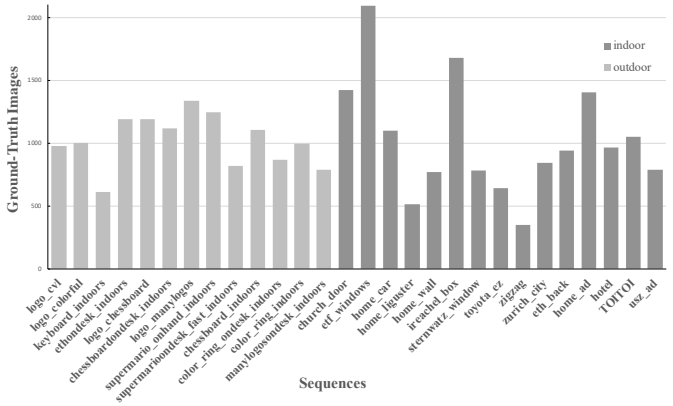


Fig. 7. The distribution of number of ground-truth images per sequence of HighREV dataset. The x-axis denotes the sequences and y-axis denotes the number of images.

4 HighREV Dataset

For event-based low-level tasks, such as event-based image deblurring and event-based frame interpolation, most works evaluate their models on datasets originally designed for image-only methods and having only synthetic events. This is because (1) event cameras are not easy to acquire yet, (2) most event cameras are of low resolution and monochrome [23], [24], [50], and (3) high-resolution chromatic datasets [53], [54] are not publicly available. To fill this gap, we record a high-quality chromatic event-image dataset for training, fine-tuning and evaluating event-based methods for frame interpolation and deblurring.

In Time Lens [53], to construct an event-based high-resolution dataset, the authors combine a synchronized, high-resolution (1280×720) event camera with an RGB camera to make a hybrid sensor. However, the alignment of the two sensors introduces error both in the temporal axis and the spatial axis. Our HighREV dataset is collected using one sensor that outputs both events and RGB frames at the same time, with a resolution of 1632×1224. Because it is a Dynamic and Active VIsion Sensor (DAVIS) [5], events and RGB images are aligned by design.

As Fig. 7 shows, our HighREV dataset consists of 30 sequences with a total of 28589 sharp images and corresponding events. We use 19934 images for training/fine-tuning and 8655 images for evaluation. The size of each RGB image is 1632×1224. The events and images are spatially aligned in the sensor. Each event has only one channel (intensity), with pixel coordinates, timestamp and polarity. 70% of the video sequences are used for training and 30% for testing and we keep the ratio of indoor and outdoor scenes approximately the same in each subset. For the collection of the dataset, the exposure time of the camera is set to 15ms and the f-stop of the lens is set to 2. The frame rate of the APS image is set to 25.

The HighREV dataset can be used for event-based sharp frame interpolation. To evaluate event-based blurry frame interpolation, we synthesize blurry images by averaging 11 consecutive original sharp frames. For blurry frame interpolation, we skip 1 or 3 sharp frames (denoted as *11+1* or *11+3* in Tab. 2). To the best of our knowledge, among all event-image datasets, our dataset has the highest resolution.

5 Experiments

5.1 Tasks and Datasets

For synthetic dataset, we use the popular GoPro dataset [31] for training and evaluation. GoPro provides blurry images, paired sharp images, and sharp image sequences used to synthesize blurry frames. The images have a size of 1280×720. We leverage the event camera simulator ESIM [37] to generate simulated event data with threshold c following a Gaussian distribution $N(\mu = 0.2, \sigma = 0.03)$. For the real-world dataset, the proposed HighREV dataset is employed. For different tasks, the datasets are as follows:

Sharp frame interpolation. The high-frame-rate sharp images of GoPro and HighREV are leveraged by skipping 7 or 15 frames and keeping the next one. The quantitative results are calculated over all the skipped frames.

Blurry frame interpolation. We synthesize blurry frames by averaging 11 sharp high-FPS frames in GoPro and HighREV. Between each blurry frame, we skip 1 or 3 frames for the evaluation of blurry frame interpolation (denoted as "11+1" or "11+3" in Tab. 2). The metrics of PSNR and SSIM are average quantitative results over all the skipped frames.

Single image deblurring. We use GoPro with synthesized blurry images (averaged from 7 or 11 sharp frames). For a real-world test, we also fine-tune and evaluate methods on REBlur [50]. We only use a single image and its corresponding events in the event branch as input, for a fair comparison.

Self-supervised training/fine-tuning. We use GoPro and HighREV for self-supervised training or fine-tuning. Experiments on both blurry frame interpolation and single-image motion deblurring are conducted. All the experiments are conducted by self-supervised training except the self-supervised fine-tuning experiment with pre-trained weights (denoted as *pre-trained ssl.* in Tab. 4). Note that the model in self-supervised learning experiments adopts fewer parameters than the model used in supervised setting.

For blurry frame interpolation and sharp frame interpolation, we train all the models on each training set and evaluate on the respective test set.

5.2 Implementation Details

Different from warping-based methods [52], [53], REFID is an end-to-end network. All its components are optimized from scratch in a single training round, without any pre-trained modules, which makes it train easier. We crop the input images and event voxels to 256×256 for training and use horizontal and vertical flips, random noise and hot pixels in event voxels [49]. Adam [25] with an initial learning rate of 2×10^{-4} and a cosine learning rate annealing strategy with 2×10^{-4} as minimum learning rate are adopted for optimization. We train the model on GoPro with a batch size of 1 for 200k iterations on 4 NVIDIA Titan RTX GPUs. For experiments on HighREV, we fine-tune the model trained on GoPro with an initial learning rate of 1×10^{-4} for 10k iterations. For image deblurring on REBlur, fine-tuning takes 600 iterations with an initial learning rate of 2×10^{-5}.

For self-supervised experiments, We consider two settings during training: (*i*) self-supervised training from scratch, where we use an initial learning rate of 1×10^{-4}, and train for $20,000$ iterations, and (*ii*) self-supervised fine-tuning with pre-trained weights, where we use an initial learning rate of 2×10^{-5} and train for $20,000$ iterations. In all settings, we use a final learning rate

TABLE 1
Comparison of sharp frame interpolation methods on GoPro and HighREV. "Frames" and "Events" indicate if a method uses frames and events for interpolation. "11+1" (resp. "11+3") indicates that the blurry image is synthesized with 11 sharp frames and 1 (resp. 3) frame(s) is skipped for frame interpolation. The number of network parameters (#Param) is also provided. *Ssl.* denotes that the model is self-supervised trained from scratch without ground-truth as supervision.

Method	Frames	Events	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	#Param
GoPro (interpolation) [31]			**7 frames skip**		**15 frames skip**		
DAIN [2]	✔	✘	28.81	0.876	24.39	0.736	24.0M
SuperSloMo [22]	✔	✘	28.98	0.875	24.38	0.747	19.8M
IFRNet [26]	✔	✘	29.84	0.920	-	-	19.7M
EDI [35]	✔	✔	18.79	0.670	17.45	0.603	0.5M
TimeReplayer [17]	✔	✔	34.02	0.960	-	-	-
Time Lens [53]	✔	✔	34.81	0.959	33.21	0.942	-
REFID	✔	✔	**36.80**	**0.980**	**35.635**	**0.974**	15.9M
HighREV (interpolation)			**7 frames skip**		**15 frames skip**		
EDI [35]	✔	✔	22.32	0.716	18.65	0.654	0.5M
RIFE [21]	✔	✘	32.28	0.904	28.22	0.864	9.8M
Time Lens [53]	✔	✔	32.81	0.901	27.06	0.810	-
REFID	✔	✔	**38.38**	**0.977**	**37.58**	**0.975**	15.9M

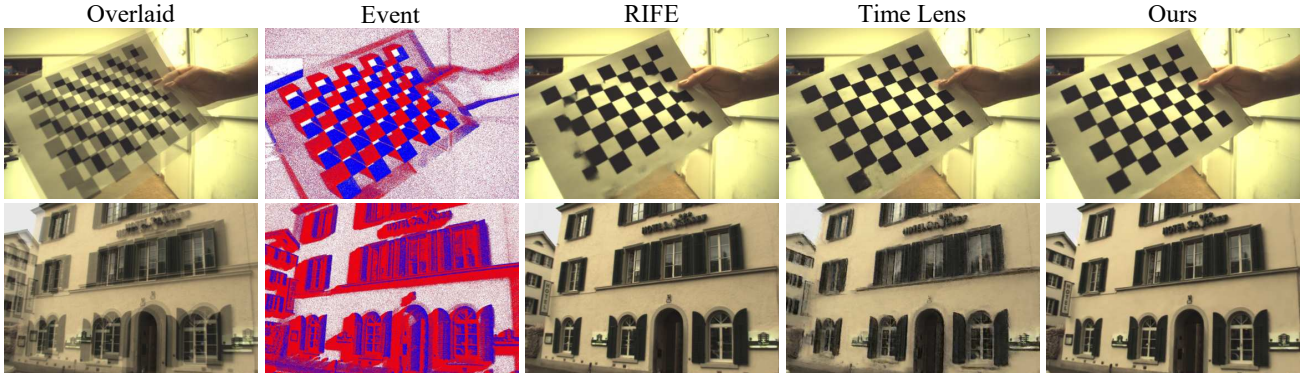


Fig. 8. **Qualitative results for sharp frame interpolation on HighREV.** RIFE [21] suffers from motion ambiguity because of the lack of event information. Time Lens [53] is vulnerable to noise. Our REFID shows superior performance both on indoor and outdoor scenes.

of 10^{-7}, and a batch size of 2. All the experiments are conducted on a single Titan RTX GPU.

5.3 Sharp Frame Interpolation

We present the results of sharp frame interpolation in Table 1. Our proposed method demonstrates state-of-the-art performance in the 7- and 15-skip settings across both examined datasets, manifesting substantial improvements over competing methods. Specifically, our approach achieves a notable enhancement of 1.99 dB and 2.43 dB in the GoPro dataset and 5.99 dB and 9.36 dB in the HighREV dataset, respectively, compared to the current state-of-the-art. Qualitative results on HighREV, illustrated in Figure 8, highlight the efficacy of our method. Notably, RIFE exhibits artifacts attributable to the ambiguity of motion between the two images, while our method maintains stable performance across diverse scenes, encompassing both indoor and outdoor environments.

Experiments on the BS-ERGB [52] dataset please refer to the Supplementary Materials.

5.4 Blurry Frame Interpolation

We compare our method with state-of-the-art image-only and event-based methods. Since most event-based methods do not have public implementations, we use "E2VID+" by adding an extra encoder for images and introduce images as extra inputs for the event-based image reconstruction method E2VID [38]. As a two-stage method, we use EFNet+IFRNet by combining a state-of-the-art event-based image deblurring method [50] with an image-only frame interpolation method [26]. For a fair comparison, IFRNet is also fed with event voxels from two directions as inputs. For Time Lens [53], because the training code is not available, we use the public model and pre-trained weights.

Quantitative results are reported in Tab. 2. Although our method can also interpolate latent frames in the exposure time, the results are reported on the interpolated frame between the two exposure times. REFID achieves 2.08 dB/0.012 and 1.29 dB/0.005 improvement in PSNR and SSIM on the "11+1" setting on GoPro and HighREV, respectively. For the "11+3" setting, the improvements over the second-best method amount to 2.08 dB/0.017 and 1.14 dB/0.005, showing that our principled bidirectional architecture with event-guided fusion leverages events more effectively. Even in the absence of ground truth for training (i.e., the *ssl.* version of REFID), our method surpasses EDI and Time Lens, underscoring the efficacy of the proposed self-supervised training framework.When applying supervised pre-training on the synthetic dataset (GoPro) and adapting to real-world dataset with the proposed self-supevisde framework, the result increases from 32.01 dB to 36.06 dB in the "11+1" setting.. The state-of-the-art event-based method Time Lens exhibits a large performance

TABLE 2
Comparison of blurry frame interpolation methods on GoPro and HighREV. Read as Tab. 1.

Method	Frames	Events	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	#Param
GoPro [31]			**11+1**		**11+3**		
REFID (*ssl.*)	✔	✔	**28.61**	**0.849**	**26.72**	**0.791**	**7.8M**
RIFE [21]	✔	✘	28.69	0.856	26.91	0.798	9.8M
EDI [35]	✔	✔	18.72	0.506	18.49	0.486	0.5M
Time Lens [53]	✔	✔	21.56	0.581	21.47	0.587	72.9M
EVDI [61]	✔	✔	29.17	0.880	28.77	0.873	0.4M
EFNet+IFRNet [26], [50]	✔	✔	33.05	0.955	32.89	0.950	28.2M
E2VID+ [39]	✔	✔	33.82	0.961	33.39	0.954	15.3M
REFID	✔	✔	**35.90**	**0.973**	**35.47**	**0.971**	15.9M
HighREV			**11+1**		**11+3**		
REFID (*ssl.*)	✔	✔	**32.01**	**0.891**	**31.16**	**0.881**	**7.8M**
RIFE [21]	✔	✘	32.79	0.904	31.24	0.890	9.8M
EDI [35]	✔	✔	24.48	0.735	23.53	0.715	0.5M
EFNet+IFRNet [26], [50]	✔	✔	35.97	0.959	35.42	0.966	28.2M
REFID (*ssl. fine-tuning*)	✔	✔	**36.05**	**0.961**	**35.51**	**0.966**	**7.8M**
E2VID+ [39]	✔	✔	36.36	0.970	35.77	0.968	15.3M
REFID	✔	✔	**37.65**	**0.975**	**36.91**	**0.973**	15.9M

TABLE 3
Comparison of single image motion deblurring methods on GoPro [31] and REBlur [50]. HINet+: event-enhanced versions of HINet [11]. *Ssl.* denotes that the model is self-supervised trained from scratch without ground-truth as supervision.

Method	Events	PSNR ↑	SSIM ↑	#Param
GoPro [31]				
EDI [35]	✔	27.34	0.901	
REFID (*ssl.*)	✔	**28.88**	**0.912**	**7.8M**
D^2Nets† [43]	✔	31.60	0.940	-
LEMD† [23]	✔	31.79	0.949	-
MPRNet [60]	✘	32.66	0.959	20.0M
Restormer [59]	✘	32.92	0.961	26.1M
ERDNet [9]	✔	32.99	0.935	-
NAFNet [10]	✘	33.69	0.967	-
EFNet [50]	✔	35.46	0.972	8.5M
REFID	✔	**35.91**	**0.973**	15.9M
REBlur [50]				
REFID (*ssl.*)	✔	**35.01**	**0.953**	**7.8M**
SRN [51]	✘	35.10	0.961	10.3M
NAFNet [10]	✘	35.48	0.962	67.9M
Restormer [59]	✘	35.50	0.959	26.1M
EDI [35]	✔	36.52	0.964	0.5M
HINet+ [11]	✔	37.68	0.973	88.9M
EFNet [50]	✔	38.12	**0.975**	8.5M
REFID	✔	**38.34**	**0.975**	15.9M

degradation on blurry frame interpolation because of the assumption of sharp keyframes and neglecting the intensity changes within the exposure time. Fig. 9 shows qualitative results. Fig. 9 (a) depicts the results for the left, right and interpolated frame on HighREV. EDI [35] is vulnerable to noise and inaccurate events. E2VID+ exhibits artifacts because its unidirectional architecture does not leverage future events. REFID achieves sharp and faithful results both on textured regions and edges thanks to the bidirectional architecture and its event-guided attention fusion. Fig.9 (b) shows the results of frame interpolation within the exposure time.

5.5 Single Image Deblurring

As a by-product, REFID can also perform single-image motion deblurring, and Tab. 3 reports quantitative comparisons on this task. It is worth mentioning that REFID predicts a short sequence of images that are within the exposure time of the input blurry image. And this is different from other traditional single-image deblurring methods. Thus, single-image deblurring results from our REFID represent the averaged PSNR and SSIM from the resulting image sequences.

With the supervised-training strategy, compared with the state-of-the-art EFNet [50], our method pushes the performance further to 35.91 dB in PSNR on GoPro. The 0.22 dB improvement in PSNR over EFNet on REBlur also evidences the robustness of REFID on real-world blurry scenes.

Considering a more realistic scene, the ground-truth sharp images for most of the event cameras are not available or hard to get. In these conditions, the models trained on synthetic data are used for real conditions. SRN+, HINet+, EFNet, and REFID in Tab.4 are examples. For our self-supervised training framework, we can choose either training from scratch on the real-world dataset (HighREV dataset) or self-supervised fine-tuning with pre-trained weights from synthetic datasets, with 1.32 dB and 3.98 dB higher than state-of-the-art.

Because of the domain gap between the synthetic data and realistic data, the performance of the models is downgraded. We choose REFID as a representative method and show the qualitative results in Fig.10. Although the result in the middle of the exposure time is good, the other results in the exposure time are vulnerable to the influence of accumulated noise. The "Left" and "Right" results in Fig.10 shows the artifacts. The settings for EDI are similar to our method because it is also a self-supervised method, but it fails to deal with the spatially-variant contrast threshold, leading to artifacts in the blurry areas. Equipped with our proposed self-supervised training framework (denoted as *ssl.* in Tab. 3), our method utilizes predicted optical flow as a substitute for events, getting rid of the influence of accumulated noise. The qualitative results show the strong generalization of the proposed framework on real-world data. Note that because our proposed self-supervised framework is built on the assumption that the model predicts a short video clip with the timestamps within the exposure time of the input image, traditional single-image deblurring methods like EFNet is not applicable to our method.

5.6 Ablation Study

Supervised training: Ablation studies on supervised training are conducted on GoPro with the "11+1" setting to analyze the ef-

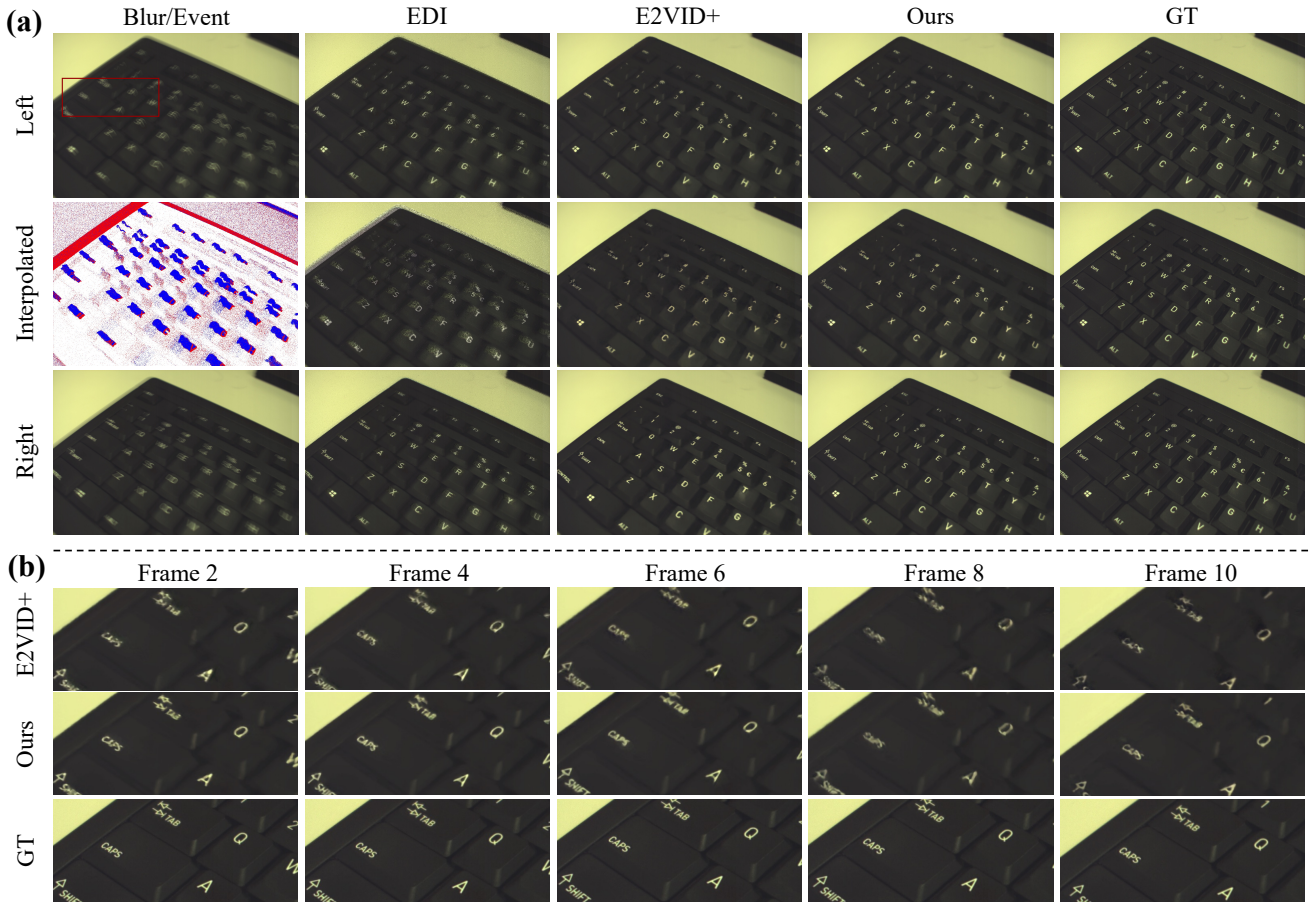


Fig. 9. (a): Visual comparison on HighREV of the restored left, right and interpolated frame. E2VID+: image-enhanced version of E2VID [38]. Compared to other event-based methods, our method achieves the most faithful results. (b): The interpolated frames in the exposure time of the left blurry image. Best viewed on a screen and zoomed in.

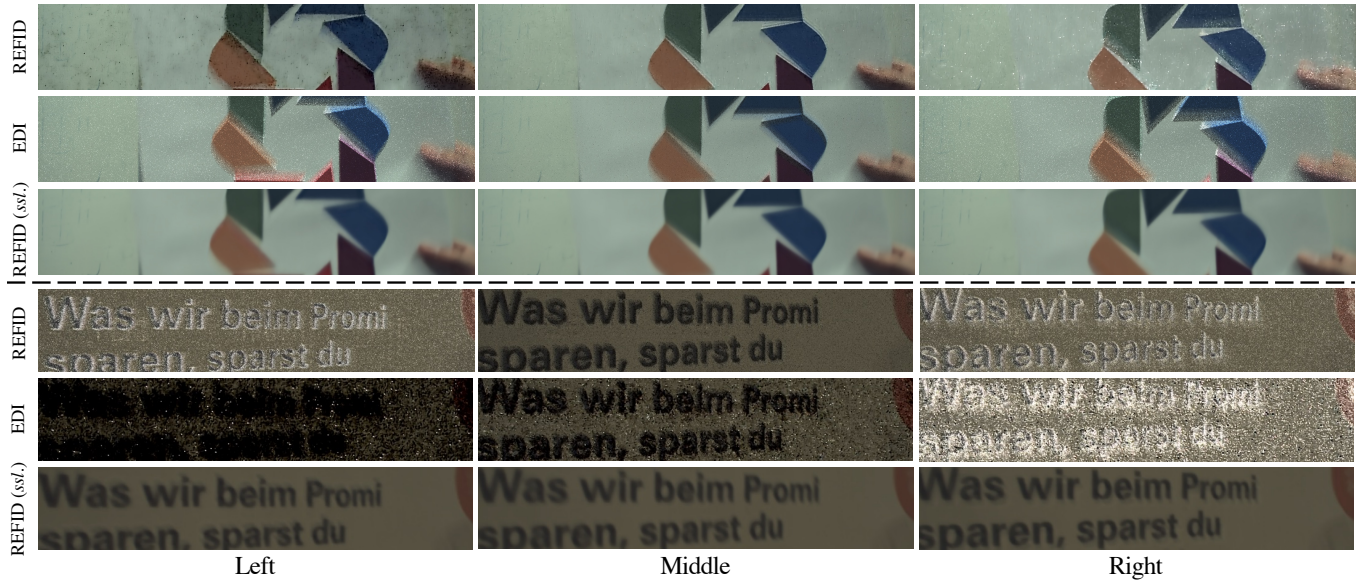


Fig. 10. **Qualitative results of the single-image deblurring without ground-truth.** The terms “Left”, “Middle”, and “Right” denote the first, middle, and last latent sharp images within the exposure time, respectively. Our method, incorporating a self-supervised framework, effectively mitigates noise accumulation in the events, ensuring consistent and high-quality deblurring results in the generated video clip.

TABLE 4
Comparison of event-based motion deblurring methods on the HighREV dataset. Methods with a + denote event-enhanced versions.

Method	sup. train. on GoPro	self sup. train. on HighREV	PSNR↑	SSIM↑
EDI [35]	✘	✔	25.32	0.753
SRN+ [51]	✔	✘	25.65	0.836
HINet+ [11]	✔	✘	28.70	0.910
EFNet [50]	✔	✘	29.55	0.936
REFID	✔	✘	28.72	0.910
REFID (*ssl.*)	✘	✔	30.04	0.931
REFID (*pre-trained ssl.*)	✔	✔	32.70	0.951

TABLE 5
Ablation study of different architectural components of our method on the GoPro [31] dataset using the "11+1" setting.

Multi-scale connection	Fusion	Recurrent	PSNR	SSIM
✘	add	✘	33.24	0.950
✔	add	✘	33.61	0.952
✔	add	ConvLSTM	34.39	0.962
✔	add	ConvGRU	34.54	0.962
✔	add	EVR unidir.	35.36	0.968
✔	add	EVR bidir.	35.81	0.971
✔	EGACA	EVR bidir.	**36.12**	**0.974**

TABLE 6
Ablation on losses. Brightness increment ($\mathcal{L}_{\text{br. inc.}}$), blur consistency ($\mathcal{L}_{\text{blur cons.}}$), and warping loss ($\mathcal{L}_{\text{warp}}$).

$\mathcal{L}_{\text{br. inc.}}$	$\mathcal{L}_{\text{blur cons.}}$	$\mathcal{L}_{\text{warp}}$	GoPro		HighREV	
			PSNR	SSIM	PSNR	SSIM
✘	✔	✘	23.52	0.824	23.41	0.817
✔	✔	✘	23.78	0.831	27.81	0.925
✔	✘	✔	7.22	0.026	10.26	0.175
✘	✔	✔	27.72	0.893	28.53	0.927
✔	✔	✔	28.88	0.912	30.04	0.931

TABLE 7
Ablation on components in self-supervised learning framework. IWE: Using images of warped events for training. Voxel Norm: the method used for voxel normalization.

IWE	Voxel Norm	GoPro		HighREV	
		PSNR	SSIM	PSNR	SSIM
✘	RobustNorm [48]	27.95	0.896	28.98	0.894
✔	RobustNorm [48]	27.99	0.895	29.25	0.924
✘	RobustNorm+	28.59	0.909	30.04	0.931
✔	RobustNorm+	**28.88**	**0.912**	**30.94**	**0.937**

fectiveness of the proposed model architecture and its components (Tab. 5). First, the proposed recurrent architecture improves PSNR by 1.75 dB compared to the non-recurrent architecture, proving the effectiveness of temporal modeling of events. Furthermore, the proposed bidirectional EVR block yields an improvement of 0.45 dB in PSNR compared to its unidirectional counterpart, showcasing the informativeness of future events and the merit of our physically-based model design. Compared to ConvLSTM [45] and ConvGRU [46], which model longer time dependencies and are used in video recognition, our EVR block using a simple ResNet [16] yields 0.84 dB improvement in PSNR. Moreover, the proposed EGACA contributes an improvement of 0.31 dB, evidencing the benefit of mining and fusing image features with adaptive weights from current events. The multi-scale connection between the image branch and the event branch also brings a 0.37 dB gain in PSNR. Finally, all our contributions together yield a substantial improvement of 2.88 dB in PSNR and 0.024 in SSIM over the baseline.

Self-supervised training: Ablation studies in Tab. 6 and Tab. 7 are conducted with self-supervised training from scratch on each dataset.

Tab. 6 shows the distinct contributions of different loss components — brightness increment loss, blur consistency loss, and warp loss — to the performance of our image deblurring model, and reports PSNR and SSIM on GoPro and HighREV. The blur consistency loss appears critical for achieving decent performance, as reflected by the results where this loss was excluded (i.e., the combination of brightness increment loss and warp loss). The model performance dropped dramatically on both datasets, reaching a PSNR of only 7.22 dB and 10.26 dB on GoPro and HighREV respectively, with correspondingly low SSIM values. On the other hand, the inclusion of the warp loss is highly beneficial. For instance, when this loss was added to the blur consistency loss, the PSNR on the HighREV dataset increased substantially from 23.41 dB to 28.53 dB, with similar improvements on the synthetic GoPro dataset. The addition of the brightness increment loss to the blur loss led to only minor improvements on the synthetic GoPro but substantially benefits the real-world results on the HighREV dataset where the PSNR increases from 23.41 dB to 27.81 dB. This difference in performance boost could stem from the fact that the events in the GoPro dataset are generated with interpolated frames that often yield subtle artifacts that affect synthetic event generation. Finally, it is worth noting that the combination of all three losses yielded the best results. The model achieved a PSNR of 28.88 dB on GoPro and 30.04 dB on HighREV, with high SSIM values, demonstrating the complementary roles of these loss components. This finding reinforces the effectiveness of our proposed framework that capitalizes on the different constraints offered by each of these losses in the image deblurring process.

Next we investigate the use of images of warped events at the input of our method, and the use of RobustNorm+, a voxel normalization method that deviates slightly from the RobustNorm in [49]. Both normalize non-zero values as

$$\begin{aligned}\bar{\mathbf{V}}_i^{\text{RN}}(\mathbf{x}) &= \text{clip}\left(\frac{\mathbf{V}_i(\mathbf{x}) - \mathbf{V}_{i,1}}{\mathbf{V}_{i,99} - \mathbf{V}_{i,1}}, 0, 1\right), \\ \bar{\mathbf{V}}_i^{\text{RN+}}(\mathbf{x}) &= \text{clip}\left(\frac{\mathbf{V}_i(\mathbf{x})}{\mathbf{V}_{i,99}}, 0, 1\right),\end{aligned} \tag{19}$$

where $\mathbf{V}_{i,p}$ denotes the p^{th} percentile of $\mathbf{V}_i(\mathbf{x})$. We found that RobustNorm+ was more stable during training and lead to fewer outliers on samples with few events. We report an ablation of these components in Tab. 7. Removing IWE from the input led to a 0.19 dB reduction on GoPro, and a 0.90 dB reduction on HighREV, justifying its use. We argue that the IWE is essential since it provides a sharp template for the network for placing edges at the output. We also find that RobustNorm+ increases the PSNR by 0.89 dB on GoPro and 1.69 dB on HighREV.

6 Conclusion

In this paper, we have considered the tasks of event-based sharp frame interpolation and blurry frame interpolation jointly, as motion blur may or may not occur in input videos depending on the speed of the motion and the length of the exposure time. To

solve these tasks with a single method, we have proposed REFID, a novel bidirectional recurrent neural network which performs fusion of the reference video frames and the corresponding event stream. The recurrent structure of REFID allows the effective propagation of event-based information across time, which is crucial for accurate interpolation. Moreover, we have introduced EGACA, a new adaptive event-image fusion module based on channel attention. In order to provide a more realistic experimental setting for the examined low-level event-based tasks, we have presented HighREV, a new event-RGB dataset with the highest spatial event resolution among related sets. We have thoroughly evaluated our network on standard event-based sharp frame interpolation, event-based blurry frame interpolation, and single-image deblurring and shown that it consistently outperforms existing state-of-the-art methods on GoPro and HighREV. To improve the generalization of the model, we propose a self-supervised training framework with warped events for blurry frame interpolation and single-image deblurring. Three loss functions are utilized for the proposed framework. Experiments on the real-world dataset are conducted to show the effectiveness of the framework. We hope the work can inspire more event-based computational imaging work for realistic applications.

In pursuit of enhancing the model's generalization, we further introduce a self-supervised training framework incorporating warped events for both blurry frame interpolation and single-image deblurring. The proposed framework integrates three distinct loss functions to constrain the model training. Through comprehensive experiments conducted on real-world datasets, we substantiate the effectiveness of our proposed approach. We envision that this work will serve as a catalyst for inspiring further exploration in the domain of event-based computational imaging for realistic applications.

REFERENCES

[1] Robert Anderson, David Gallup, Jonathan T Barron, Janne Kontkanen, Noah Snavely, Carlos Hernández, Sameer Agarwal, and Steven M Seitz. Jump: virtual reality video. *ACM Transactions on Graphics (TOG)*, 35(6):1–13, 2016. 1

[2] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proc. CVPR*, pages 3703–3712, 2019. 8

[3] Yuhan Bao, Lei Sun, Yuqin Ma, Diyang Gu, and Kaiwei Wang. Improving fast auto-focus with event polarity. *arXiv preprint arXiv:2303.08611*, 2023. 3

[4] José M Bioucas-Dias, Mario AT Figueiredo, and Joao Pedro Oliveira. Total variation-based image deconvolution: a majorization-minimization approach. In *ICASSP*, volume 2, pages II–II. IEEE, 2006. 3

[5] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 1, 7

[6] Samuel Bryner, Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. Event-based, direct camera tracking from a photometric 3d map using nonlinear optimization. In *Proc. ICRA*, pages 325–331. IEEE, 2019. 6

[7] Ayan Chakrabarti, Todd Zickler, and William T Freeman. Analyzing spatially-varying blur. In *Proc. CVPR*, pages 2512–2519. IEEE, 2010. 3

[8] Huaijin Chen, Jinwei Gu, Orazio Gallo, Ming-Yu Liu, Ashok Veeraraghavan, and Jan Kautz. Reblur2deblur: Deblurring videos via self-supervised learning. In *Proc. ICCP*, pages 1–9. IEEE, 2018. 3

[9] Haoyu Chen, Minggui Teng, Boxin Shi, Yizhou Wang, and Tiejun Huang. Learning to deblur and generate high frame rate video with an event camera. *arXiv preprint arXiv:2003.00847*, 2020. 1, 4, 9

[10] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022. 9

[11] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. HINet: Half instance normalization network for image restoration. In *Proc. CVPRW*, 2021. 9, 11

[12] Shengyang Dai and Ying Wu. Motion from blur. In *Proc. ICCV*, pages 1–8. IEEE, 2008. 3

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

[14] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2020. 1

[15] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Eklt: Asynchronous photometric feature tracking using events and frames. *International Journal of Computer Vision*, 128(3):601–618, 2020. 6

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016. 11

[17] Weihua He, Kaichao You, Zhendong Qiao, Xu Jia, Ziyang Zhang, Wenhui Wang, Huchuan Lu, Yaoyuan Wang, and Jianxing Liao. Timereplayer: Unlocking the potential of event cameras for video interpolation. In *Proc. CVPR*, pages 17804–17813, 2022. 1, 2, 3, 4, 5, 8

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. CVPR*, pages 7132–7141, 2018. 2, 6

[19] Ping Hu, Simon Niklaus, Stan Sclaroff, and Kate Saenko. Many-to-many splatting for efficient video frame interpolation. In *Proc. CVPR*, pages 3553–3562, 2022. 1

[20] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proc. CVPR*, pages 1312–1321, 2021. 2

[21] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. RIFE: Real-time intermediate flow estimation for video frame interpolation. *arXiv preprint arXiv:2011.06294*, 2020. 2, 8, 9

[22] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proc. CVPR*, pages 9000–9008, 2018. 1, 8

[23] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning event-based motion deblurring. In *Proc. CVPR*, pages 3320–3329, 2020. 2, 4, 7, 9

[24] Taewoo Kim, Jungmin Lee, Lin Wang, and Kuk-Jin Yoon. Event-guided deblurring of unknown exposure time videos. *arXiv preprint arXiv:2112.06988*, 2021. 2, 7

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[26] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang. Ifrnet: Intermediate feature refine network for efficient frame interpolation. In *Proc. CVPR*, pages 1969–1978, 2022. 2, 8, 9

[27] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. *ACM Transactions on Graphics (TOG)*, 26(3):70–es, 2007. 3

[28] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Understanding and evaluating blind deconvolution algorithms. In *Proc. ICCV*, pages 1964–1971. IEEE, 2009. 3

[29] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning event-driven video deblurring and interpolation. In *Proc. ECCV*, pages 695–710. Springer, 2020. 1, 4

[30] Peidong Liu, Joel Janai, Marc Pollefeys, Torsten Sattler, and Andreas Geiger. Self-supervised linear motion deblurring. *IEEE Robotics and Automation Letters*, 5(2):2475–2482, 2020. 3

[31] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proc. CVPR*, pages 3883–3891, 2017. 2, 7, 8, 9, 11

[32] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *Proc. CVPR*, pages 5437–5446, 2020. 1

[33] Jihyong Oh and Munchurl Kim. Demfi: Deep joint deblurring and multi-frame interpolation with flow-guided attentive correlation and recursive boosting. *arXiv preprint arXiv:2111.09985*, 2021. 2

[34] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. Deblurring face images with exemplars. In *Proc. ECCV*, pages 47–62. Springer, 2014. 3

[35] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proc. CVPR*, pages 6820–6829, 2019. 1, 2, 4, 8, 9, 11

[36] Federico Paredes-Vallés and Guido CHE de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proc. CVPR*, pages 3446–3455, 2021. 3, 6

[37] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on robot learning*, pages 969–982. PMLR, 2018. 2, 7

[38] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proc. CVPR*, pages 3857–3866, 2019. 4, 8, 10

[39] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proc. CVPR*, 2019. 4, 9

[40] Wenqi Ren, Jingang Zhang, Xiangyu Xu, Lin Ma, Xiaochun Cao, Gaofeng Meng, and Wei Liu. Deep video dehazing with semantic segmentation. *IEEE Transactions on Image Processing*, 28(4):1895–1908, 2018. 1

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 2015. 4

[42] Qi Shan, Jiaya Jia, and Aseem Agarwala. High-quality motion deblurring from a single image. *ACM Transactions on Graphics (TOG)*, 27(3):1–10, 2008. 3

[43] Wei Shang, Dongwei Ren, Dongqing Zou, Jimmy S Ren, Ping Luo, and Wangmeng Zuo. Bringing events into video deblurring with non-consecutively blurry frames. In *Proc. ICCV*, pages 4531–4540, 2021. 9

[44] Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao. Blurry video frame interpolation. In *Proc. CVPR*, pages 5114–5123, 2020. 2

[45] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *NIPS*, 28, 2015. 5, 11

[46] Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. *NIPS*, 30, 2017. 5, 11

[47] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow. In *Proc. ECCV*, pages 628–645. Springer, 2022. 4, 6

[48] Timo Stoffregen and Lindsay Kleeman. Event cameras, contrast maximization and reward functions: An analysis. In *Proc. CVPR*, pages 12300–12308, 2019. 11

[49] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *Proc. ECCV*, pages 534–549. Springer, 2020. 7, 11

[50] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *Proc. ECCV*, pages 412–428. Springer, 2022. 2, 4, 6, 7, 8, 9, 11

[51] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proc. CVPR*, pages 8174–8182, 2018. 9, 11

[52] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proc. CVPR*, pages 17755–17764, 2022. 1, 2, 3, 4, 5, 7, 8

[53] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proc. CVPR*, pages 16155–16164, 2021. 1, 2, 3, 4, 5, 7, 8, 9

[54] Patricia Vitoria, Stamatios Georgoulis, Stepan Tulyakov, Alfredo Bochicchio, Julius Erbach, and Yuanyou Li. Event-based image deblurring with dynamic motion awareness. *arXiv preprint arXiv:2208.11398*, 2022. 7

[55] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proc. CVPR*, pages 10081–10090, 2019. 3

[56] Chao-Yuan Wu, Nayan Singhal, and Philipp Krahenbuhl. Video compression through image interpolation. In *Proc. ECCV*, pages 416–431, 2018. 1

[57] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. *NIPS*, 32, 2019. 1

[58] Lei Yu, Bishan Wang, Xiang Zhang, Haijian Zhang, Wen Yang, Jianzhuang Liu, and Gui-Song Xia. Learning to super-resolve blurry images with events. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[59] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proc. CVPR*, pages 5728–5739, 2022. 9

[60] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proc. CVPR*, pages 14821–14831, 2021. 9

[61] Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *Proc. CVPR*, pages 17765–17774, 2022. 3, 4, 9

[62] C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM Transactions on Graphics (TOG)*, 23(3):600–608, 2004. 1

[63] Liad Pollak Zuckerman, Eyal Naor, George Pisha, Shai Bagon, and Michal Irani. Across scales and across dimensions: Temporal super-resolution using deep internal learning. In *Proc. ECCV*, pages 52–68. Springer, 2020. 2

Lei Sun received his B.S. degree in Optical Engineering from Beijing Institute of Technology (BIT) in 2018, He is currently pursuing his Ph.D. degree from the National Research Center for Optical Instrumentation, Zhejiang University (ZJU), where he was awarded Xiaomi Special Scholarship and CSC Scholarship. He was a visiting doctoral student at Computer Vision Lab, ETH Zürich. He was also a visiting doctoral student at Robotics and Perception Group, University of Zürich. His research interests include semantic segmentation, event-based vision, and low-level vision. For more information, visit his website: https://ahupujr.github.io/.

Daniel Gehrig obtained his Ph.D. with the highest distinction from the Robotics and Perception Group (RPG), at the University of Zürich, Switzerland for which he was granted the UZH Annual Award. Between 2012 to 2018 he completed his master's studies in Mechanical Engineering at ETH Zürich, with the highest possible score, for which he was awarded the Willi Studer Prize and ETH Medal for the best master's thesis of the year. His research interests lie at the intersection of robotics, computer vision, and machine learning for event-based vision. His work has been featured prominently in IEEE Spectrum and on popular channels like Two Minute Papers.

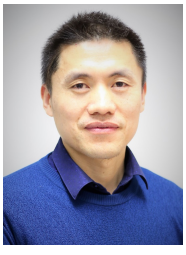
Christos Sakaridis is a lecturer at ETH Zürich and a senior postdoctoral researcher at the Computer Vision Lab of ETH Zürich. His research fields are computer vision and machine learning. The focus of his research is on semantic and geometric visual perception, involving multiple domains, visual conditions, and visual or non-visual modalities. Since 2021, he is the Principal Engineer in TRACE-Zürich, a large-scale project on computer vision for autonomous cars and robots. He received the ETH Zürich Career Seed Award in 2022. He obtained his PhD from ETH Zürich in 2021, having worked in Computer Vision Lab. Prior to that, he received his MSc in Computer Science from ETH Zürich in 2016 and his Diploma in Electrical and Computer Engineering from National Technical University of Athens in 2014.

Mathias Gehrig obtained his M.Sc. in Robotics, Systems and Control from ETH Zürich, Switzerland, in 2016, after receiving his B.Sc. in Mechanical Engineering in 2013. As a Ph.D. candidate in computer science at the University of Zürich, supervised by Prof. Davide Scaramuzza, he focuses on the application of machine learning for real-time computer vision and robotics. Notably, his work was nominated for the Best Paper Award at the 2023 Conference on Computer Vision and Pattern Recognition (CVPR).

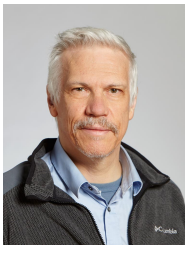
Jingyun Liang is currently a PhD student at Computer Vision Lab, ETH Zürich. He received his B.S degree and master's degree from National University of Defense Technology in 2014 and 2016, His research focuses on low-level vision research, especially on image and video restoration, such as super-resolution, deblurring and denoising.

Peng Sun received his B.S degree in Optical Engineering from Zhejiang University(ZJU) in 2020, and gain his master degree in information engineering in 2023 from the National Research Center, Zhejiang University. His research focuses on event-based vision, image deblurring, and visual detection algorithms.

Zhijie Xu , received his PhD from the University of Derby, UK, in 2000. He is a senior academic director of Graduate Education and director of the Centre for Visual and Immersive Computing at the University of Huddersfield He has published 200 peer-reviewed papers and edited five books. He has mentored 60 postgraduates, including 15 PhD students to completion. He has served as General Chair for conferences, including the 23rd IEEE ICAC.



Kaiwei Wang is currently a full professor at the State Key Laboratory of Modern Optical Instrumentation, and the Deputy Director of the National Optical Instrument Engineering Research Center at Zhejiang University. He received a B.S. degree in 2001 and a Ph.D. degree in 2005 respectively, both from Tsinghua University. In October 2005, he started his postdoctoral research at the Center of Precision Technologies (CPT) of Huddersfield University, funded by the Royal Society International Visiting Postdoctoral Fellowship and the British Engineering Physics Council. He joined Zhejiang University in February 2009 and has been mainly researching on intelligent optical sensing technology and visual assisting technology for the visually impaired. Up to date, he owns 80 patents and has published more than 150 refereed research papers. For more information, visit his Website: http://wangkaiwei.org/.



Luc Van Gool is both a full professor with KU Leuven (Belgium) and with ETH Zürich (Switzerland). His main area of expertise is computer vision. He received the Koenderink Prize with the European Conference on Computer Vision, in 2016, the David Marr prize (best paper award) at the International Conference on Computer Vision, in 1998 and the U.V. Helava Award, one of the most prestigious ISPRS awards, in 2012. He was also awarded an ERC Advanced Grant, in 2011 for his project VarCity (Variation & the City), was nominated 'Distinguished Researcher' by the IEEE Society of Computer Vision, in 2017, and received the 5-yearly excellence prize by the Flemish Fund for Scientific Research, in 2016. He received several other best paper awards as well. He is co-founder of several spin-offs. He has been involved in the organization of several, major conferences and as an associate editor for multiple, first-tier scientific journals.

Davide Scaramuzza is a Professor of Robotics and Perception at the University of Zurich. He did his Ph.D. at ETH Zurich, a postdoc at the University of Pennsylvania, and was a visiting professor at Stanford University. His research focuses on autonomous, agile microdrone navigation using standard and event-based cameras. He pioneered autonomous, vision-based navigation of drones, which inspired the navigation algorithm of the NASA Mars helicopter and many drone companies. He contributed significantly to visual-inertial state estimation, vision-based agile navigation of microdrones, and low-latency, robust perception with event cameras, which were transferred to many products, from drones to automobiles, cameras, AR/VR headsets, and mobile devices. In 2022, his team demonstrated that an AI-controlled, vision-based drone could outperform the world champions of drone racing, a result that was published in Nature. He is a consultant for the United Nations on disaster response and disarmament. He has won many awards, including an IEEE Technical Field Award, the IEEE Robotics and Automation Society Early Career Award, a European Research Council Consolidator Grant, a Google Research Award, two NASA TechBrief Awards, and many paper awards. In 2015, he co-founded Zurich-Eye, today Meta Zurich, which developed the world-leading virtual-reality headset Meta Quest. In 2020, he co-founded SUIND, which builds autonomous drones for precision agriculture. Many aspects of his research have been featured in the media, such as The New York Times, The Economist, and Forbes.