Generalizing Event-Based Motion Deblurring in Real-World Scenarios

Xiang Zhang[1], Lei Yu[1✉], Wen Yang[1], Jianzhuang Liu[2], Gui-Song Xia[1]

[1]Wuhan University   [2]Shenzhen Institute of Advanced Technology

{xiangz, ly.wd, yangwen, guisong.xia}@whu.edu.cn, jz.liu@siat.ac.cn

Abstract

Event-based motion deblurring has shown promising results by exploiting low-latency events. However, current approaches are limited in their practical usage, as they assume the same spatial resolution of inputs and specific blurriness distributions. This work addresses these limitations and aims to generalize the performance of event-based deblurring in real-world scenarios. We propose a scale-aware network that allows flexible input spatial scales and enables learning from different temporal scales of motion blur. A two-stage self-supervised learning scheme is then developed to fit real-world data distribution. By utilizing the relativity of blurriness, our approach efficiently ensures the restored brightness and structure of latent images and further generalizes deblurring performance to handle varying spatial and temporal scales of motion blur in a self-distillation manner. Our method is extensively evaluated, demonstrating remarkable performance, and we also introduce a real-world dataset consisting of multi-scale blurry frames and events to facilitate research in event-based deblurring.

Multimedia Material

The Multi-Scale Real-world Blurry Dataset (MS-RBD) and our Pytorch implementation are available at: `https://github.com/XiangZ-0/GEM`.

1. Introduction

Due to the fixed exposure time of frame-based cameras, motion blur often occurs in scenes with dynamic targets or camera ego-motion, degrading the quality of the acquired images [15, 33]. Conventional motion deblurring approaches attempt to resolve this by exploiting deconvolution and blur kernel estimation techniques [30, 16], and recent research further improves the deblurring performance with the advanced deep-learning methods [12, 36]. However, tra-
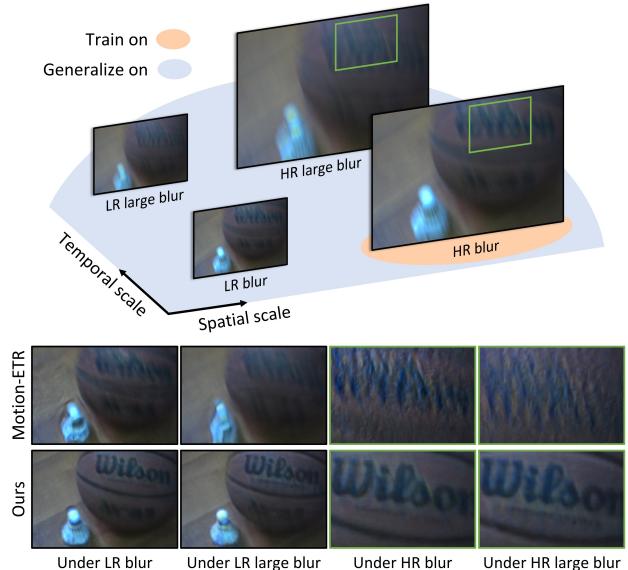
Figure 1: An illustrative example of motion deblurring via the state-of-the-art algorithm Motion-ETR [36] and our proposed method, which is trained on HR blurry frames and LR events in a self-supervised manner and can generalize to the inputs at different temporal and spatial scales.

ditional frame-based methods usually assume specific motion patterns, *e.g.*, linear or quadratic motion trajectory, for blurry images and thus often face challenges in real-world scenarios with complex non-uniform motions. In addition, due to the motion ambiguity and texture erasure issues in blurry images [29, 35], frame-based approaches often struggle to extract the precise motion and restore the accurate latent images from severely blurred frames.

The advent of event cameras poses a paradigm shift in visual perception and information acquisition, benefiting a wide variety of applications [17, 6, 22, 34, 10, 37, 32, 7, 31]. For motion deblurring tasks, the microsecond-level low latency of events enables almost continuous observation of dynamic scenes and alleviates the motion ambiguity in blurry frames [21, 26]. Moreover, the brightness

changes recorded in event streams inherently correspond to high-contrast edges, compensating for the intensity texture erased by motion blur [29, 35, 24, 13]. However, the performance of current event-based deblurring methods is usually confined to the distribution of training data, *e.g.*, frames with a certain range of blurriness and the same spatial resolution as events, posing limitations in real-world scenarios.

- **Temporal Limitation:** Most previous approaches synthesize or collect blurry frames in a fixed range of exposure time for training [26, 29], which implicitly assumes motion blur with a specific distribution of blurriness. However, real-world motion blur often violates this assumption in highly dynamic scenes, resulting in a performance drop of pre-trained models.

- **Spatial Limitation:** Existing methods mainly take frames and events of the same spatial resolution as input, ignoring that frame-based cameras usually have larger spatial resolution than event-based ones in practice [6]. Besides, due to the varying distributions of events at different spatial scales [9], how to effectively deblur High-Resolution (HR) frames with Low-Resolution (LR) events remains an open problem.

In this paper, we propose to address the above issues and generalize the performance of event-based motion deblurring in both spatial and temporal domains, as shown in Fig. 1. In detail, a Scale-Aware Network (SAN) is first designed to extract high frame-rate HR sequences from a single HR blurry frame and its concurrent LR events. Inspired by implicit neural representation [3], we implement a Multi-Scale Feature Fusion (MSFF) module to represent frame and event features in a spatially continuous manner, which allows flexible setups of input spatial resolutions. In the temporal dimension, an Exposure-Guided Event Representation (EGER) is presented to enable the arbitrary selection of target latent images without requiring model modification or re-training. To fit real-world data distribution, a two-stage self-supervised learning framework is further proposed. In the first stage, we efficiently supervise the restored brightness and structure of latent images by utilizing the relativity of blurriness. Following that, a self-distillation strategy is applied to generalize the deblurring performance to handle varying spatial and temporal scales of motion blur. Overall, our contributions are three-fold:

- A scale-aware network is presented to allow flexible setups of input spatial resolutions and output temporal scales, which is able to restore high frame-rate HR sequences from HR blurry frames and LR events.

- A two-stage self-supervised learning framework is proposed to efficiently fit real-world data distributions and generalize deblurring performance to handle varying spatial and temporal scales of motion blur.

- A real-world dataset MS-RBD containing HR blurry frames and LR events is built to facilitate deblurring research. Extensive experiments on both synthetic and real datasets validate the effectiveness of our approach.

2. Related Work

Motion Deblurring. How to recover sharp images from motion-blurred frames has been investigated for decades [5, 20, 30, 16, 12, 36, 15, 33]. Conventional deblurring methods often model the blurred image as a latent sharp image convolved with a blur kernel in the presence of additive noise [5], and several techniques have been adopted for motion deblurring, including deconvolution [16], kernel estimation [30], and dark channel prior [20]. Recently, deep-learning approaches are also employed to achieve better deblurring results and extract video sequences from blurry frames [12, 36]. By exploiting an ordering-invariant constraint, LEVS gradually resumes the temporal ordering embedded in motion blur and recovers sharp sequences from a blurry input [12]. Motion-ETR further improves deblurring performance by utilizing Deformable Convolutional Networks (DCNs) [38] to predict the motion trajectory within blurry frames, which tackles temporal disorder and enables the recovery of non-linear exposure trajectories [36].

However, traditional frame-based methods usually assume specific motion patterns of blurry frames and thus often fail in real-world scenarios with complex non-uniform motion. Besides, large motion blur will eliminate the intensity texture in the acquired frames, posing challenges to recovering satisfied latent images from blurry inputs.

Event-based Motion Deblurring. Recent works have revealed the advantages of events in motion deblurring [21, 26, 29, 23, 35, 24, 13, 31]. With the low latency and high temporal resolution of event cameras [6], events naturally encode the information of high-contrast texture and precise motion of dynamic scenes, facilitating the reconstruction of sharp latent images under complex motion. Previous work of [21] first establishes the Event-based Double Integral (EDI) model for motion deblurring, which bridges the blurry frames and latent sharp images with events. Following that, learning-based methods are developed to achieve better results by adopting techniques like sparse coding [26, 31], parametric polynomial [23], and cross-modal attention [24]. To fit real-world data distribution, recent works also focus on learning from real blurry frames and events by semi-/self-supervised methods [29, 35].

Although event-based methods have made significant progress in motion deblurring, the aforementioned approaches generally focus on deblurring frames with specific temporal scales of motion blur and the same spatial resolutions as events, showing limitations in real-world applications. In our approach, a scale-aware network is designed to deblur HR frames with LR events and simultaneously en-

able flexible setups of input spatial resolutions. Moreover, a self-supervised learning framework is proposed to efficiently fit real-world data distribution and generalize the deblurring performance in both spatial and temporal domains.

3. Method

In this section, we first formulate event-based deblurring and our goal in Sec. 3.1. Based on this, we then introduce the scale-aware network in Sec. 3.2 and finally propose our self-supervised learning method in Sec. 3.3.

3.1. Problem Formulation

We first review the basic model of event-based motion deblurring, which aims to restore sharp latent images from blurry frames and the corresponding events. According to the event generation model [17, 6], each event is emitted asynchronously whenever the log-scale brightness change reaches the event threshold $c > 0$,

$$\log(I(t, \mathbf{x})) - \log(I(f, \mathbf{x})) = p \cdot c, \tag{1}$$

where $\log(I(t, \mathbf{x})), \log(I(f, \mathbf{x}))$ correspond to the log-scale intensity of pixel $\mathbf{x}$ at time t and f, and $p \in \{+1, -1\}$ denotes the polarity showing the direction of brightness change. On the other hand, blurry frames can be formulated as the average of the latent images within the exposure period $\mathcal{T}$ [2] (pixel position $\mathbf{x}$ is omitted for readability),

$$B_T = \frac{1}{T} \int_{t \in \mathcal{T}} I(t) dt, \tag{2}$$

where B_T indicates the blurry frame captured with exposure time T. Combining Eq. (1) and (2), one can bridge blurry frames and sharp images by the EDI model [21],

$$I(t) = \frac{B_T}{E(t, \mathcal{T})}, \text{ with} \tag{3}$$

$$E(t, \mathcal{T}) = \frac{1}{T} \int_{f \in \mathcal{T}} \exp(c \int_t^f e(s) ds) df, \tag{4}$$

where $e(\tau) \triangleq p \cdot \delta(\tau - t)$ indicates the continuous event representation and $\delta(\cdot)$ denotes the Dirac function. Since directly restoring $I(t)$ via Eq. (3) often suffers from the instability of event threshold c in practice [6, 29], learning-based approaches are employed to better fit the statistics of events [26, 29], which are generally in the form of

$$I(t) = \text{Deblur}(t; B_T, \mathcal{E}_\mathcal{T}), \quad \forall t \in \mathcal{T}, \tag{5}$$

where $\text{Deblur}(\cdot)$ denotes a motion deblurring network and $\mathcal{E}_\mathcal{T}$ indicates the events triggered within $\mathcal{T}$.

Define the spatial resolution ratio of frames to events as $\mathcal{R}(B_T, \mathcal{E}_\mathcal{T})$, e.g., $\mathcal{R}(B_T, \mathcal{E}_\mathcal{T}) = 4$ means the resolution of

frame B_T is four times that of events $\mathcal{E}_\mathcal{T}$, previous learning-based approaches are commonly trained on the dataset

$$\mathcal{D}(\mathbf{T}, \mathbf{R}) \triangleq \{B_T, \mathcal{E}_\mathcal{T} | T \in \mathbf{T}, \ \mathcal{R}(B_T, \mathcal{E}_\mathcal{T}) \in \mathbf{R}\} \tag{6}$$

with $\mathbf{R} = \{1\}$ indicating the same spatial resolution of frames and events, and $\mathbf{T} = \{T_k\}_{k=1}^K$ denoting a set composed of K exposure parameters. Once trained, it is difficult to directly apply previous methods to process real-world inputs with $\mathcal{R}(B_T, \mathcal{E}_\mathcal{T}) > 1$, i.e., HR blurry frames and LR events. Besides, the set $\mathbf{T}$ implicitly assumes a specific distribution of blurriness, which often results in a performance drop of pre-trained models when inferring more blurred frames.

To foster the application of event-based motion deblurring in real-world scenarios, it is necessary to enlarge the sets of $\mathbf{T}$ and $\mathbf{R}$. However, collecting sufficient datasets to cover a wide range of $\mathbf{T}$, $\mathbf{R}$ is time-consuming and impractical. Also, sharp ground-truth images are difficult to collect when recording real-world blurry datasets and thus are usually unavailable for training. Therefore, the goal of our work is to design a fully self-supervised deblurring algorithm that only needs to train on a dataset $\mathcal{D}(\mathbf{T}, \{\bar{R}\})$ with $\forall \bar{R} \geq 1$ to fit real-world setups, but is able to generalize on a larger set $\mathcal{D}(\mathbf{T}^*(M), \mathbf{R}^*(\bar{R}))$ as shown in Fig. 1, where

$$\mathbf{T}^*(M) \triangleq \sum_{m=1}^M \{mT_k\}_{k=1}^K,$$
$$\mathbf{R}^*(\bar{R}) \triangleq \{R | 1 \leq R \leq \bar{R}, R \in \mathbb{R}\}, \tag{7}$$

with $M \in \mathbb{N}^+$ denoting a parameter that can be chosen to determine the temporal scale of motion blur.

3.2. Scale-Aware Network
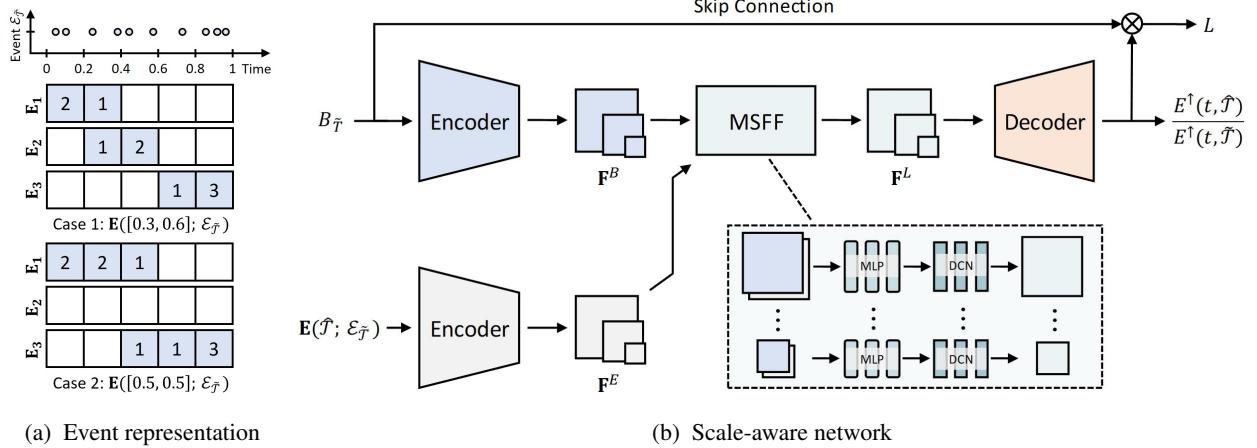
Unlike previous methods that focus on fitting Eq. (3), our Scale-Aware Network (SAN) aims to approximate a more general function to allow flexible input spatial scales and enable learning from different temporal scales of motion blur. Due to the different spatial scales of frames and events in our task, we first modify Eq. (3) to

$$I(t) = \frac{B_T}{E^\uparrow(t, \mathcal{T})}, \tag{8}$$

with $E^\uparrow(t, \mathcal{T})$ indicating the upsampled version of $E(t, \mathcal{T})$ to match the spatial resolution of B_T. Next we consider a more blurred frame $B_{\tilde{T}}$ and similarly get $I(t) = B_{\tilde{T}}/E^\uparrow(t, \tilde{\mathcal{T}})$ with $T < \tilde{T}$ and $\mathcal{T} \subset \tilde{\mathcal{T}}$. For the same target image $I(t)$, one can then derive

$$B_T = \frac{E^\uparrow(t, \mathcal{T})}{E^\uparrow(t, \tilde{\mathcal{T}})} B_{\tilde{T}}, \tag{9}$$

which converts the more blurred frame $B_{\tilde{T}}$ into its less blurred latent image B_T. Inspired by this, we design our

Figure 2: (a) An example of our Exposure-Guided Event Representation (EGER). The event stream $\mathcal{E}_{\tilde{\mathcal{T}}}$ contains 10 negative events in $\tilde{\mathcal{T}} = [0, 1]$. We show two cases of $\mathbf{E}(\hat{\mathcal{T}}; \mathcal{E}_{\tilde{\mathcal{T}}})$ under $N = 5$ with $\hat{\mathcal{T}} = [0.3, 0.6]$ and $\hat{\mathcal{T}} = [0.5, 0.5]$, and the operation is the same for positive events. (b) Structure of our proposed network with the Multi-Scale Feature Fusion (MSFF) module.

SAN to approximate a general function

$$L = \frac{E^{\uparrow}(t, \hat{\mathcal{T}})}{E^{\uparrow}(t, \tilde{\mathcal{T}})} B_{\tilde{T}} \approx \text{SAN}(\hat{\mathcal{T}}; B_{\tilde{T}}, \mathcal{E}_{\tilde{\mathcal{T}}}), \qquad (10)$$

where $\hat{\mathcal{T}} \subset \tilde{\mathcal{T}}$ controls the output temporal scale of the target latent image L. Thus, our SAN is able to restore both sharp and blurry latent images by setting different $\hat{\mathcal{T}}$, *i.e.*,

- **Blur2sharp conversion:** If $\hat{\mathcal{T}} = [t, t]$, $E^{\uparrow}(t, \hat{\mathcal{T}}) = 1$ holds since no event is integrated, and thus $L = I(t)$.

- **Blur2blur conversion:** If $\hat{\mathcal{T}} = \mathcal{T}$, the target function in Eq. (10) becomes Eq. (9), and thus $L = B_T$.

This enables SAN to learn from blur2blur conversion without requiring sharp ground-truth images. Moreover, our SAN does not assume the same spatial resolution of inputs. To fulfill the temporal and spatial flexibility, an Exposure-Guided Event Representation (EGER) and a Multi-Scale Feature Fusion (MSFF) module are respectively proposed.
Exposure-Guided Event Representation. The goal of EGER is to explicitly model the conversion relationship between the input blurry frame and the latent image with events, which can be regarded as preparing events for computing $E^{\uparrow}(t, \hat{\mathcal{T}})/E^{\uparrow}(t, \tilde{\mathcal{T}})$ in Eq. (10). Given an event stream $\mathcal{E}_{\tilde{\mathcal{T}}}$ with $\tilde{\mathcal{T}} \triangleq [t_s, t_e]$ and the target exposure period $\hat{\mathcal{T}} \triangleq [\hat{t}_s, \hat{t}_e]$, we first evenly divide $\tilde{\mathcal{T}}$ into N temporal bins and generate three $2N \times H \times W$ event tensors $\mathbf{E}_1, \mathbf{E}_2$, and $\mathbf{E}_3$ with $2, H, W$ indicating event polarity, height, and width. The three tensors $\mathbf{E}_1, \mathbf{E}_2$, and $\mathbf{E}_3$ accumulate the events split based on the intervals $[t_s, \hat{t}_s], [\hat{t}_s, \hat{t}_e]$, and $[\hat{t}_e, t_e]$, respectively. By simple event splitting, $\mathbf{E}_2$ contains events $\mathcal{E}_{\hat{\mathcal{T}}}$ in the target exposure period for comput-

ing $E^{\uparrow}(t, \hat{\mathcal{T}})$, and the combination of $\mathbf{E}_1, \mathbf{E}_2$, and $\mathbf{E}_3$ corresponds to events $\mathcal{E}_{\tilde{\mathcal{T}}}$ for $E^{\uparrow}(t, \tilde{\mathcal{T}})$. Then our EGER is formed by concatenating the three event tensors, *i.e.*,

$$\mathbf{E}(\hat{\mathcal{T}}; \mathcal{E}_{\tilde{\mathcal{T}}}) = \text{Concat}(\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3), \qquad (11)$$

where $\mathbf{E}(\hat{\mathcal{T}}; \mathcal{E}_{\tilde{\mathcal{T}}})$ is the EGER of target exposure period $\hat{\mathcal{T}}$ conditioned on the input events $\mathcal{E}_{\tilde{\mathcal{T}}}$.

As the toy example shown in Fig. 2a, the input event stream can be represented as different $\mathbf{E}(\hat{\mathcal{T}}; \mathcal{E}_{\tilde{\mathcal{T}}})$ according to the chosen $\hat{\mathcal{T}}$. This allows SAN to determine the output temporal scales and recover both blurry (*e.g.*, case 1 in Fig. 2a) and sharp (*e.g.*, case 2 in Fig. 2a) latent images from the same input. Also, EGER enables flexible selection of $\hat{\mathcal{T}} \subset \tilde{\mathcal{T}}$ for arbitrarily high frame-rate video generation.
Multi-Scale Feature Fusion. Another challenge for SAN is the different spatial resolutions between HR blurry frames and LR events. Inspired by the Local Implicit Image Function (LIIF) [3] that represents images in a spatially continuous manner, we propose to fuse frames and events by learning a continuous feature representation. As depicted in Fig. 2b, we first extract multi-scale blur and event features $\mathbf{F}^B = \{F_i^B\}, \mathbf{F}^E = \{F_i^E\}$ with F_i^B, F_i^E denoting the features at the i-th scale by two encoder networks (our encoder and decoder networks are split from an hourglass network, detailed in the supplementary material). Considering the cross-sensor gap between frame-based and event-based cameras, we use the blur features to provide brightness reference and guide the upsampling of event features in our MSFF module. Specifically, a Multi-Layer Perceptron (MLP) is employed to predict the fused feature value from cross-modal local features, *i.e.*, $f_i(z) = \text{MLP}_i(z, s; F_i^B, F_i^E)$, where z indicates a 2D coordinate in the continuous spatial domain, $f_i(z)$ is the pre-

dicted feature at z, and $s = [s_h, s_w]$ is the size of the target feature pixel. Afterward, the coarsely fused feature is refined by a DCN with a larger receptive field, generating the final feature of latent image $F_i^L = \text{DCN}_i(f_i)$. We finally pass the features $\mathbf{F}^L$ through a decoder network and restore the latent image L.

By utilizing the MSFF module, our SAN is able to effectively fuse the information of frames and events at different spatial resolutions. In addition, since the coordinates are continuous, MSFF enables flexible setups of input spatial scales, *e.g.*, our SAN can simultaneously take inputs of $\mathcal{R}(B_T, \mathcal{E}_\mathcal{T}) = 4$ and $\mathcal{R}(B_T, \mathcal{E}_\mathcal{T}) = 2.5$ without network modification or re-training, facilitating practical usage.

3.3. Self-Supervised Learning

Our self-supervised learning approach consists of two stages: we first constrain the restored brightness and structure of latent images by utilizing the relativity of blurriness, and then generalize the deblurring performance in both temporal and spatial dimensions via self-distillation techniques. **Brightness and Structure Consistency.** Based on Eq. (10), we propose to constrain the reconstruction brightness by learning blur2blur conversion. Given B_T from a blurry video, we synthesize a more blurred image $B_{\tilde{T}}$ by averaging M adjacent blurry frames of B_T ($M = 2$ in our experiments) and formulate the constraint as

$$\mathcal{L}_{BC} = \|B_T - \text{SAN}(\mathcal{T}; B_{\tilde{T}}, \mathcal{E}_{\tilde{\mathcal{T}}})\|_1, \qquad (12)$$

which efficiently ensures brightness consistency by learning to restore B_T from $B_{\tilde{T}}$.

According to Eq. (8), recovering the structure of sharp latent images is equivalent to estimating accurate $E^\uparrow(t, \mathcal{T})$ for each $I(t)$. To achieve this, we first breakdown our SAN into $\text{SAN}(\hat{\mathcal{T}}; B_{\tilde{T}}, \mathcal{E}_{\tilde{\mathcal{T}}}) = \text{SAN}^E(\hat{\mathcal{T}}; B_{\tilde{T}}, \mathcal{E}_{\tilde{\mathcal{T}}}) \cdot B_{\tilde{T}}$, where $\text{SAN}^E(\cdot)$ estimates the event ratio based on Eq. (10) and Fig. 2b,

$$\frac{E^\uparrow(t, \hat{\mathcal{T}})}{E^\uparrow(t, \tilde{\mathcal{T}})} \approx \text{SAN}^E(\hat{\mathcal{T}}; B_{\tilde{T}}, \mathcal{E}_{\tilde{\mathcal{T}}}). \qquad (13)$$

By setting $\hat{\mathcal{T}} = [t, t]$, $\text{SAN}^E(\cdot)$ is able to estimate $E^\uparrow(t, \tilde{\mathcal{T}})$ for restoring sharp latent image $I(t)$, *i.e.*,

$$\frac{1}{E^\uparrow(t, \tilde{\mathcal{T}})} \approx \text{SAN}^E([t, t]; B_{\tilde{T}}, \mathcal{E}_{\tilde{\mathcal{T}}}). \qquad (14)$$

Then, we constrain the structure of the restored $I(t)$ by

$$\mathcal{L}_{SC} = \left\| \text{SAN}^E(\mathcal{T}; B_{\tilde{T}}, \mathcal{E}_{\tilde{\mathcal{T}}}) - \frac{\text{SAN}^E([t, t]; B_{\tilde{T}}, \mathcal{E}_{\tilde{\mathcal{T}}})}{\text{SAN}^E([t, t]; B_T, \mathcal{E}_\mathcal{T})} \right\|_1, \qquad (15)$$

where $\text{SAN}^E(\mathcal{T}; B_{\tilde{T}}, \mathcal{E}_{\tilde{\mathcal{T}}})$ provides strong supervision to avoid collapsing solutions as it is constrained in $\mathcal{L}_{BC}$. $\mathcal{L}_{SC}$

guarantees structure recovery by transferring the knowledge learned from blur2blur to the blur2sharp case. With $\mathcal{L}_{BC}$ and $\mathcal{L}_{SC}$, SAN efficiently achieves motion deblurring by ensuring the brightness and structure of sharp latent images. **Temporal and Spatial Generalization.** The second stage of training aims to generalize the deblurring performance of SAN in both temporal and spatial dimensions. For temporal generalization, we propose a self-distillation loss

$$\mathcal{L}_{TG} = \|\overline{\text{SAN}}([t, t]; B_T, \mathcal{E}_\mathcal{T}) - \text{SAN}([t, t]; B_{\tilde{T}}, \mathcal{E}_{\tilde{\mathcal{T}}})\|_1, \qquad (16)$$

where $\overline{\text{SAN}}$ indicates a fixed teacher model pre-trained using $\mathcal{L}_{BC}$ and $\mathcal{L}_{SC}$, and SAN is the student network loaded from $\overline{\text{SAN}}$ and continuing to train. Since $\overline{\text{SAN}}$ can recover relatively more reliable latent images from the less blurred frame B_T, we treat the output of $\overline{\text{SAN}}$ as pseudo-ground-truth images and teach SAN to deblur the more blurred frame $B_{\tilde{T}}$, which improves the deblurring ability of SAN and generalizes its performance to handle different temporal scales of motion blur.

With the above constraints, our SAN learns to deblur HR frames with LR events at a fixed spatial ratio $\mathcal{R}(B_T, \mathcal{E}_\mathcal{T}) = \bar{R}$, but its performance in handling different spatial scales of motion blur, *i.e.*, $\mathcal{R}(B_T, \mathcal{E}_\mathcal{T}) \in \mathbf{R}^*(\bar{R})$, is not guaranteed. To generalize the deblurring performance in the spatial domain, we encourage SAN to adaptively project event features according to the input spatial scale $\mathcal{R}(B_T, \mathcal{E}_\mathcal{T})$. Specifically, we first form inputs with varying spatial scales by randomly down-sampling $B_{\tilde{T}}$ to $B_{\tilde{T}}^\downarrow$ with $\forall \mathcal{R}(B_{\tilde{T}}^\downarrow, \mathcal{E}_{\tilde{\mathcal{T}}}) \in [1, \bar{R}]$, and then formulate the constraint based on the idea of self-distillation,

$$\mathcal{L}_{SG} = \|\overline{\text{SAN}}^\downarrow([t, t]; B_T, \mathcal{E}_\mathcal{T}) - \text{SAN}([t, t]; B_{\tilde{T}}^\downarrow, \mathcal{E}_{\tilde{\mathcal{T}}})\|_1, \qquad (17)$$

where $\overline{\text{SAN}}^\downarrow$ means $\overline{\text{SAN}}$ followed by a down-sampling operation. With $\mathcal{L}_{SG}$, our SAN is able to propagate the deblurring performance under $\mathcal{R}(B_T, \mathcal{E}_\mathcal{T}) = \bar{R}$ to different input spatial scales $\mathcal{R}(B_T, \mathcal{E}_\mathcal{T}) \in \mathbf{R}^*(\bar{R})$.

Finally, our self-supervised learning framework can be summarized as

$$\mathcal{L} = \beta_{BC}\mathcal{L}_{BC} + \beta_{SC}\mathcal{L}_{SC} + \beta_{TG}\mathcal{L}_{TG} + \beta_{SG}\mathcal{L}_{SG}, \qquad (18)$$

where $\beta_{BC}, \beta_{SC}, \beta_{TG}, \beta_{SG}$ indicate the balancing parameters. Compared with the previous self-supervised EVDI [35], our method generalizes the deblurring performance to handle the varying blurriness levels and different spatial scales of real motion blur. Furthermore, our approach shows better efficiency by design. For example, EVDI supervises brightness consistency via reblurring techniques, which require restoring a large number (49 in EVDI) of latent images per input during training, while ours efficiently fulfills this by learning blur2blur conversion (please see the supplementary material for detailed comparisons).
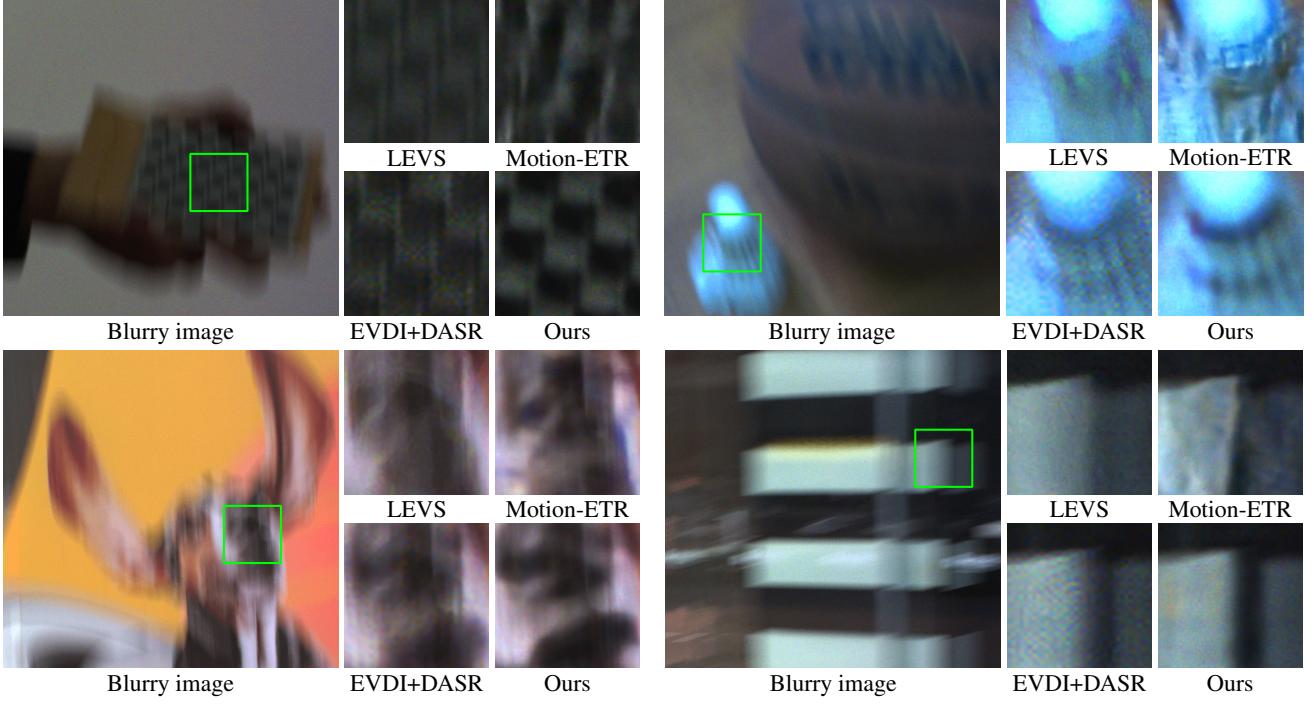
Figure 3: Qualitative comparisons under real-world HR frames and LR events on our MS-RBD.
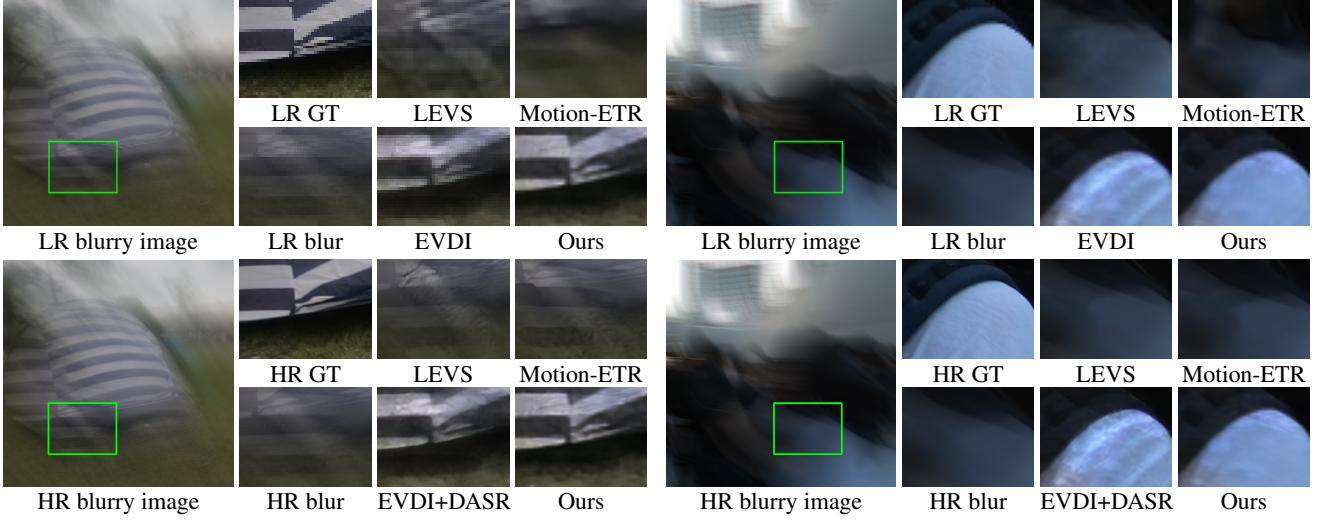

Figure 4: Qualitative comparisons under different spatial scales $\mathcal{R}(B_T, \mathcal{E}_T) = 1$ (LR blur, top row) and $\mathcal{R}(B_T, \mathcal{E}_T) = 4$ (HR blur, bottom row) on the Ev-REDS dataset. GT indicates ground-truth images.

4. Experiments and Analysis

4.1. Experimental Setup

Datasets. Three different datasets containing synthetic, semi-synthetic, and real-world blurry frames and events are employed in our experiments for evaluation.

Ev-REDS: We build a synthetic dataset upon REDS [19] for evaluation on different spatial scales. We first crop the sharp images to size 1280×640 and down-sample them to 320×160 to form HR and LR sequences. For each sequence, we generate high frame-rate videos by interpolating 7 images between consecutive frames using RIFE [11], and then synthesize blurry frames by averaging 49 sharp images of the high frame-rate videos. Events are generated via VID2E [8] on the LR sequences to form two sets with different spatial scales $\mathcal{R}(B_T, \mathcal{E}_T) = 4$ (HR frames and LR

Table 1: Quantitative comparisons under different spatial scales ($\mathcal{R}(B_T, \mathcal{E}_T) = 1$ and $\mathcal{R}(B_T, \mathcal{E}_T) = 4$) on the Ev-REDS dataset. Image (DASR), video (RealBasicVSR), and event (EventZoom) super-resolution techniques are employed to assist event-based deblurring methods in the case of $\mathcal{R}(B_T, \mathcal{E}_T) = 4$. Symbol / denotes unavailable metrics as some methods only work with gray images. For those that work with color images (LEVS, Motion-ETR, EVDI, and ours), their results are also converted to gray-scale for computing gray metrics. Best and second-best results are **bolded** and underlined, respectively.

| Method | Comparison under $\mathcal{R}(B_T, \mathcal{E}_T) = 1$ | | | | Comparison under $\mathcal{R}(B_T, \mathcal{E}_T) = 4$ | | | |
| | Color metric | | Gray metric | | Color metric | | Gray metric | |
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
LEVS [12]	18.24	0.4665	18.36	0.4680	18.62	0.4612	18.75	0.4644
Motion-ETR [36]	17.79	0.4376	17.90	0.4388	18.23	0.4292	18.34	0.4320
EDI [21] (+DASR [27])	/	/	20.41	0.6067	/	/	18.81	0.4553
eSL-Net [26]	/	/	19.41	0.7119	/	/	18.96	0.5604
RED [29] (+DASR [27])	/	/	23.21	0.7959	/	/	22.60	0.6350
EVDI [35] (+EventZoom [4])	23.88	0.7789	24.37	0.7917	18.93	0.4815	19.07	0.4848
EVDI [35] (+RealBasicVSR [1])	23.88	0.7789	24.37	0.7917	23.33	0.6441	23.79	0.6568
EVDI [35] (+DASR [27])	23.88	0.7789	24.37	0.7917	23.35	0.6368	23.83	0.6477
Ours	**24.12**	**0.7898**	**24.63**	**0.8022**	**23.95**	**0.6647**	**24.43**	**0.6749**
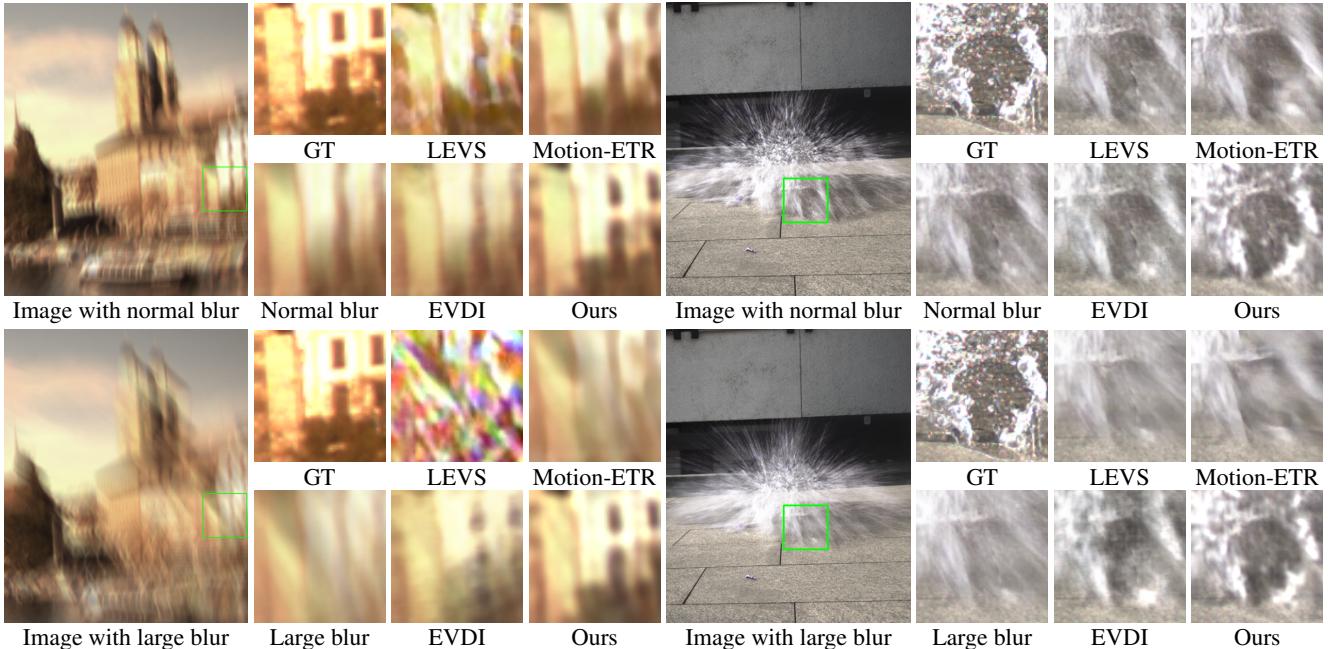
Figure 5: Qualitative comparisons under normal blur (top row) and large blur (bottom row) on the HS-ERGB dataset.

events, for training and testing) and $\mathcal{R}(B_T, \mathcal{E}_T) = 1$ (LR frames and events, only for testing).

HS-ERGB: HS-ERGB dataset [25] contains sharp videos and real events at the same spatial resolution, and thus we employ it for evaluation on different temporal scales of motion blur. We first increase the frame rate of the original videos by interpolating 7 images between consecutive frames with Time Lens [25], and then synthesize

two types of blurry videos by averaging 49 and 97 frames, which we call normal and large blur, respectively. The set with normal blur is used for training and testing, and the one with large blur is only used for testing.

MS-RBD: Due to the lack of available real-world datasets with HR blurry frames and LR events, we construct a Multi-Scale Real-world Blurry Dataset (MS-RBD) with a FLIR Blackfly S global shutter RGB camera and a

Table 2: Quantitative comparisons under $\mathcal{R}(B_T, \mathcal{E}_\mathcal{T}) = 1$ and different temporal scales (normal and large blur) on the HS-ERGB dataset. The symbol / denotes unavailable metrics as some algorithms only work with gray images. The results of color models (LEVS, Motion-ETR, EVDI, and ours) are converted to gray-scale for computing gray metrics.

Method	Comparison under normal blur				Comparison under large blur			
	Color metric		Gray metric		Color metric		Gray metric	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
LEVS [12]	22.13	0.5548	22.70	0.5935	21.72	0.5429	22.06	0.5741
Motion-ETR [36]	23.79	0.6276	24.05	0.6464	22.73	0.5842	22.88	0.6010
EDI [21]	/	/	23.93	0.7043	/	/	22.33	0.6517
eSL-Net [26]	/	/	24.10	0.6811	/	/	22.76	0.6248
RED [29]	/	/	<u>26.05</u>	0.7234	/	/	<u>24.81</u>	0.6676
EVDI [35]	<u>25.13</u>	<u>0.7072</u>	25.49	<u>0.7312</u>	<u>24.08</u>	<u>0.6637</u>	24.35	<u>0.6856</u>
Ours	**26.22**	**0.7292**	**26.87**	**0.7529**	**25.41**	**0.6936**	**25.94**	**0.7168**

DAVIS346 camera. A beam splitter is implemented in front of the two cameras with 50% splitting. In total, we collect 32 sequences of data composed of 22 indoor and 10 outdoor scenes, where the blur caused by camera ego-motion and dynamic scenes are both considered. We also set the frame rate of the FLIR camera to 30 and 15 FPS to imitate the blur at different temporal scales. After spatial alignment, each sequence contains 60 RGB frames at size 1152×768 and the corresponding 288×192 events. More details can be found in the supplementary material.

Implementation Details. Our SAN is implemented in the Pytorch platform and trained on NVIDIA GeForce RTX 2080 Ti GPUs with batch size 3. We set the number of temporal bins $N = 16$ and the temporal scale parameter $M = 2$. The Adam optimizer [14] and the SGDR schedule [18] are employed for training. We first train an SAN with the parameters $[\beta_{BC}, \beta_{SC}, \beta_{TG}, \beta_{SG}] = [50, 1, 0, 0]$ and learning rate 1×10^{-3} for 210 epochs. With the pre-trained $\overline{\text{SAN}}$ as the teacher model, we continue training the SAN with $[\beta_{BC}, \beta_{SC}, \beta_{TG}, \beta_{SG}] = [50, 1, 50, 50]$ and learning rate 5×10^{-4} for 15 cycles. Every cycle lasts for 30 epochs, and we update the teacher model at the end of each cycle.

4.2. Benchmarking

We evaluate the proposed method by comparing with the state-of-the-art deblurring approaches, including frame-based algorithms LEVS [12], Motion-ETR [36], and event-based methods EDI [21], eSL-Net [26], RED [29], and EVDI [35]. Since we assume real-world scenarios without available ground-truth images, only the self-supervised EVDI can be trained under such circumstances, and we use the official codes for re-training. In the case of $\mathcal{R}(B_T, \mathcal{E}_\mathcal{T}) > 1$, we employ state-of-the-art image, video, and event super-resolution techniques DASR [27], RealBasicVSR [1], and EventZoom [4] to assist event-

based deblurring methods as they only accept inputs of $\mathcal{R}(B_T, \mathcal{E}_\mathcal{T}) = 1$. Metrics PSNR and SSIM [28] are computed based on sequence restoration, *i.e.*, restoring 7 sharp images from one blurry input, for quantitative evaluation.

Tab. 1 validates the robust performance of our proposed approach under different spatial scales. Although frame-based methods can directly process blurry frames at different spatial resolutions, they often fail in highly dynamic scenes with complex motions because of motion ambiguity, as depicted in Fig. 3. For event-based algorithms, eSL-Net is able to produce HR results by simultaneously considering motion deblurring and image super-resolution. However, eSL-Net only receives blurry frames and events of the same spatial resolution, which limits its performance due to the information loss caused by image down-sampling. Similar to eSL-Net, previous event-based methods generally assume $\mathcal{R}(B_T, \mathcal{E}_\mathcal{T}) = 1$ for inputs, and thus super-resolution techniques are necessary to restore HR results. As shown in Tab. 1 and Fig. 4, such cascaded scheme often leads to sub-optimal performance as the deblurring or super-resolution errors will be propagated to the subsequent stage.

Regarding the case with different temporal scales, previous approaches are often limited by the blur distribution of training data, resulting in a significant performance drop when encountering large motion blur, as shown in Tab. 2 and Fig. 5. Benefiting from the temporal generalization technique in our learning framework, the proposed method can recover reliable latent images of the target scenes under both normal and large blur as depicted in Fig. 5. Thus, our method not only enables flexible setups of input spatial resolution but also exhibits promising performance in handling motion blur of different temporal scales, facilitating applications in real-world scenarios.

Table 3: Ablation study of our self-supervised learning framework under different spatial scales $\mathcal{R}(B_T, \mathcal{E}_{\mathcal{T}}) = 1$ (LR) and $\mathcal{R}(B_T, \mathcal{E}_{\mathcal{T}}) = 4$ (HR) on the Ev-REDS dataset.
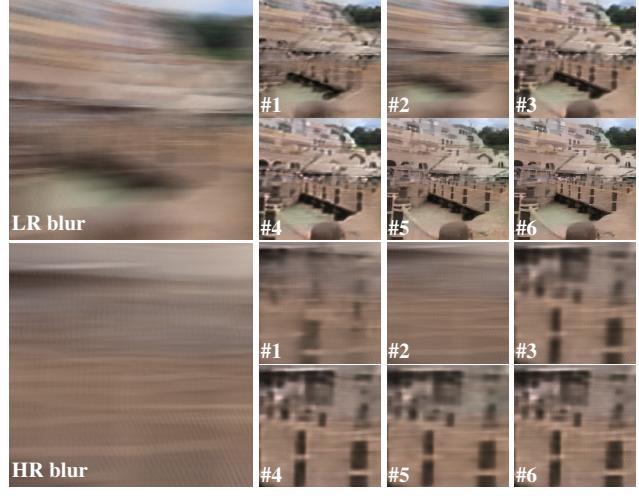
| ID	$\mathcal{L}_{BC}$	$\mathcal{L}_{SC}$	$\mathcal{L}_{TG}$	$\mathcal{L}_{SG}$	LR / HR PSNR ↑
#1	✓				19.20 / 19.95
#2		✓			18.94 / 18.95
#3	✓	✓			21.77 / 23.39
#4	✓	✓	✓		21.73 / **24.00**
#5	✓	✓		✓	<u>23.46</u> / 23.23
#6	✓	✓	✓	✓	**24.12** / <u>23.95</u>

4.3. Ablation Study
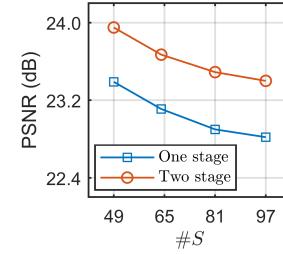
We study the contribution of each component in our self-supervised learning method on the Ev-REDS dataset and draw the following conclusions:

Combination of Brightness and Structure Consistency. In Tab. 3 and Fig. 6a, model #1 trained only with $\mathcal{L}_{BC}$ effectively constrains the restored brightness by learning blur2blur conversion, but it suffers from missing structure and produces blurry results. By combining $\mathcal{L}_{BC}$ and $\mathcal{L}_{SC}$, model #3 successfully recovers the correct structure with accurate brightness as shown in Fig. 6a, simultaneously guaranteeing brightness and structure consistency. However, training solely with $\mathcal{L}_{SC}$ will lead to collapsing solutions as the results of model #2 depicted in Fig. 6a. This is because the supervision signal in $\mathcal{L}_{SC}$ is strongly dependent on $\mathcal{L}_{BC}$ as discussed in Sec. 3.3, and thus the structure constraint $\mathcal{L}_{SC}$ should be used together with the brightness constraint $\mathcal{L}_{BC}$ to achieve motion deblurring.
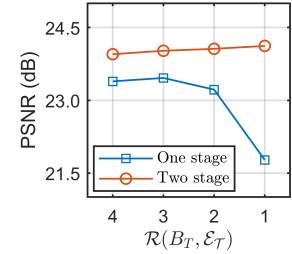
Effectiveness of Temporal and Spatial Generalization. Although the first stage of training achieves promising performance in deblurring HR blurry frames with LR events, i.e., $\mathcal{R}(B_T, \mathcal{E}_{\mathcal{T}}) = 4$, it struggles to handle different temporal and spatial scales of motion blur as shown in Fig. 6b and 6c. To improve the deblurring performance in the temporal dimension, our $\mathcal{L}_{TG}$ supervises the consistency of latent images restored from blurry frames with different levels of blurriness. Since it is generally easier to deblur the frames with normal blur (B_T) than that with large blur ($B_{\tilde{T}}$), $\mathcal{L}_{TG}$ encourages our model to produce similar results when dealing with both cases and thus learns to tackle more severe motion blur, leading to better deblurring performance (models #3 and #4 in Tab. 3 and Fig. 6a) and general improvements in large blur removal (Fig. 6b). In the spatial domain, the performance inconsistency shown in Tab. 3 and Fig. 6c is because model #3 only learns to project events to fit HR frames but neglects the varying event distributions at different spatial scales. With $\mathcal{L}_{SG}$, our SAN can adaptively adjust the learned event distribution according to

(a) Qualitative comparisons

(b) Temporal comparisons   (c) Spatial comparisons

Figure 6: Results of different models in Tab. 3 on the Ev-REDS dataset. (a) Qualitative comparisons under $\mathcal{R}(B_T, \mathcal{E}_{\mathcal{T}}) = 1$ (LR blur, top row) and $\mathcal{R}(B_T, \mathcal{E}_{\mathcal{T}}) = 4$ (HR blur, bottom row). (b, c) Comparisons of models using one-stage and two-stage training, i.e., model #3 and #6, under different temporal and spatial scales of motion blur. $\#S$ denotes the number of sharp images used to synthesize one blurry frame, and larger $\#S$ indicates more blurred frames.

$\mathcal{R}(B_T, \mathcal{E}_{\mathcal{T}})$ and propagate the deblurring performance under $\mathcal{R}(B_T, \mathcal{E}_{\mathcal{T}}) = 4$ to other spatial scales, leading to consistent performance as shown in Fig. 6a and 6c.

5. Conclusion

This paper proposes to generalize event-based motion deblurring in real-world scenarios. We first present a scale-aware network to allow flexible setups of input spatial resolutions and enable learning from different temporal scales of motion blur. Following that, a two-stage self-supervised learning framework is designed for model training with real data and performance generalization in both spatial and temporal domains. In addition, a real-world dataset containing high-resolution blurry frames and low-resolution events is released to facilitate the evaluation of frame/event-based deblurring approaches in real-world scenes.

References

[1] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *CVPR*, pages 5962–5971, 2022.

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

[3] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, pages 8628–8638, 2021.

[4] Peiqi Duan, Zihao W Wang, Xinyu Zhou, Yi Ma, and Boxin Shi. Eventzoom: Learning to denoise and super resolve neuromorphic events. In *CVPR*, pages 12824–12833, 2021.

[5] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T. Roweis, and William T. Freeman. Removing camera shake from a single photograph. *ACM Trans. Graph.*, 25(3):787–794, 2006.

[6] 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 TPAMI*, 44(1):154–180, 2020.

[7] Yue Gao, Siqi Li, Yipeng Li, Yandong Guo, and Qionghai Dai. Superfast: 200× video frame interpolation via event camera. *IEEE TPAMI*, 2022.

[8] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *CVPR*, pages 3586–3595, 2020.

[9] Daniel Gehrig and Davide Scaramuzza. Are high-resolution event cameras really needed? *arXiv preprint arXiv:2203.14672*, 2022.

[10] Javier Hidalgo-Carrió, Guillermo Gallego, and Davide Scaramuzza. Event-aided direct sparse odometry. In *CVPR*, pages 5781–5790, 2022.

[11] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *ECCV*, pages 624–642, 2022.

[12] Meiguang Jin, Givi Meishvili, and Paolo Favaro. Learning to extract a video sequence from a single motion-blurred image. In *CVPR*, pages 6334–6342, 2018.

[13] Taewoo Kim, Jeongmin Lee, Lin Wang, and Kuk-Jin Yoon. Event-guided deblurring of unknown exposure time videos. In *ECCV*, pages 519–538. Springer, 2022.

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

[15] Jaihyun Koh, Jangho Lee, and Sungroh Yoon. Single-image deblurring with neural networks: A comparative survey. *Computer Vision and Image Understanding*, 203:103134, 2021.

[16] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR*, pages 233–240, 2011.

[17] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE J. Solid-State Circuits*, 43(2):566–576, 2008.

[18] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *ICLR*, 2017.

[19] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPRW*, pages 1974–1984, 2019.

[20] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *CVPR*, pages 1628–1636, 2016.

[21] 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 *CVPR*, pages 6820–6829, 2019.

[22] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE TPAMI*, 43(6):1964–1980, 2019.

[23] Chen Song, Qixing Huang, and Chandrajit Bajaj. E-cir: Event-enhanced continuous intensity recovery. In *CVPR*, pages 7803–7812, 2022.

[24] 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 *ECCV*, pages 412–428. Springer, 2022.

[25] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *CVPR*, pages 16155–16164, 2021.

[26] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *ECCV*, pages 155–171, 2020.

[27] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *CVPR*, pages 10581–10590, 2021.

[28] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multi-scale structural similarity for image quality assessment. In *IEEE Asilomar Conf. Sign. Syst. Comput.*, volume 2, pages 1398–1402, 2003.

[29] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblurring with real events. In *ICCV*, pages 2583–2592, 2021.

[30] Li Xu and Jiaya Jia. Two-phase kernel estimation for robust motion deblurring. In *ECCV*, pages 157–170, 2010.

[31] 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 TPAMI*, 45(8):10027–10043, 2023.

[32] Lei Yu, Xiang Zhang, Wei Liao, Wen Yang, and Gui-Song Xia. Learning to see through with events. *IEEE TPAMI*, 45(7):8660–8678, 2023.

[33] Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Björn Stenger, Ming-Hsuan Yang, and Hongdong Li. Deep image deblurring: A survey. *IJCV*, 130(9):2103–2130, 2022.

[34] Xiang Zhang, Wei Liao, Lei Yu, Wen Yang, and Gui-Song Xia. Event-based synthetic aperture imaging with a hybrid network. In *CVPR*, pages 14235–14244, 2021.

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

[36] Youjian Zhang, Chaoyue Wang, Stephen J Maybank, and Dacheng Tao. Exposure trajectory recovery from motion blur. *IEEE TPAMI*, 44(11):7490–7504, 2021.

[37] Zelin Zhang, Anthony Yezzi, and Guillermo Gallego. Formulating event-based image reconstruction as a linear inverse problem with deep regularization using optical flow. *IEEE TPAMI*, (01):1–18, 2022.

[38] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, pages 9308–9316, 2019.