Learning Scale-Aware Spatio-temporal Implicit Representation for Event-based Motion Deblurring

Wei Yu [1] **Jianing Li** [2] **Shengping Zhang** [1] **Xiangyang Ji** [3]

Abstract

Existing event-based motion deblurring methods mostly focus on restoring images with the same spatial and temporal scales as events. However, the unknown scales of images and events in the real world pose great challenges and have rarely been explored. To address this gap, we propose a novel Scale-Aware Spatio-temporal Network (SASNet) to flexibly restore blurred images with event streams at arbitrary scales. The core idea is to implicitly aggregate both spatial and temporal correspondence features of images and events to generalize at continuous scales. To restore highly blurred local areas, we develop a Spatial Implicit Representation Module (SIRM) to aggregate spatial correlation at any resolution through event encoding sampling. To tackle global motion blur, a Temporal Implicit Representation Module (TIRM) is presented to learn temporal correlation via temporal shift operations with long-term aggregation. Additionally, we build a High-resolution Hybrid Deblur (H2D) dataset using a new-generation hybrid event-based sensor, which comprises images with naturally spatially aligned and temporally synchronized events at various scales. Experiments demonstrate that our SASNet outperforms state-of-the-art methods on both synthetic GoPro and real H2D datasets, especially in high-speed motion scenarios. Code and dataset are available at https://github.com/aipixel/SASNet.

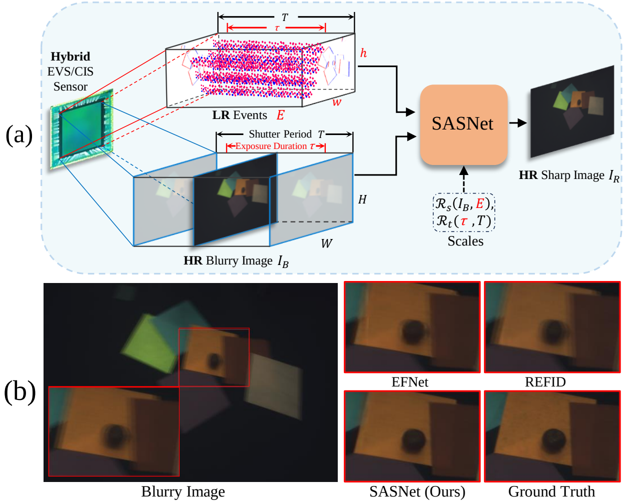


Figure 1. (a) Illustration of our arbitrary-scale event-based deblurring task, where $\mathcal{R}_s$ and $\mathcal{R}_t$ denote spatial and temporal scales, respectively. (b) Visual comparison results between EFNet (Sun et al., 2022), REFID (Sun et al., 2023), and our SASNet under the scale differences scenario.

1. Introduction

With the wide use of CMOS Image Sensors (CIS) and bio-inspired Event Vision Sensors (EVS) in robots (Rebecq et al., 2018) and autonomous vehicles (Gallego et al., 2020), event-based vision tasks have attracted increasing attention (Hidalgo-Carrió et al., 2022; Gao et al., 2022; Zhang et al., 2022; Yu et al., 2023b). Due to the microsecond-level low latency of EVS, necessary motion information can be inherently captured from events to help alleviate motion blurring in CIS frames (Chen et al., 2024b). Consequently, Event-based Motion Deblurring (EMD) has become a prominent topic in computational photography.

Despite the advancements in existing EMD solutions (Sun et al., 2023; Zhang et al., 2023b), practical applications are still hindered by two limitations. **Dataset Limitation:** Existing event-based deblurring datasets (Zhang et al., 2023a; Cho et al., 2023) are usually collected from cameras with low spatial resolution (e.g., DAVIS346 with 346×260) or binocular camera systems using beam splitters, which are

[1] School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China [2] School of Computer Science, Peking University, Beijing, China. [3] Department of Automation, Tsinghua University, Beijing, China.. Correspondence to: Shengping Zhang <s.zhang@hit.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

cumbersome and inaccurate due to the artificial spatial alignment and time synchronization of CIS and EVS. **Algorithm Limitation:** Current algorithms (Sun et al., 2022; 2023) always assume that the inputs of CIS images and EVS events have the same spatial (i.e., resolution) and temporal (i.e., exposure duration) scales, which are confined by the scale differences of different shooting equipment and environments in practice, as shown in Fig. 1 (b). Notably, capturing both high spatial and temporal resolution events is difficult in practice due to the bandwidth limitations and high costs of EVS cameras (Kim et al., 2022; Yang & Yamac, 2022).

To address these limitations, we first investigate arbitrary-scale EMD to restore High-Resolution (HR) sharp images from HR blurry images and Low-Resolution (LR) events with different spatial and temporal scales, as illustrated in Fig. 1 (a). We observe that the large-scale differences between events and images lead to limited motion deblurring effects due to insufficient utilization of the spatio-temporal corresponding feature, and there is an urgent need for a generalized model to enable deblurring at arbitrary scale differences. Based on this motivation, we propose a novel Scale-Aware Spatio-temporal Network (SASNet) to flexibly restore blurred images with event streams at arbitrary scales. The core idea is to implicitly aggregate both spatial and temporal correspondence features of images and events to generalize at continuous scales, rather than the previous explicit discrete methods (Sun et al., 2022; Chen et al., 2020; Dosovitskiy et al., 2015) at a single fixed scale. In particular, we exploit two physical characteristics of blurry scenes: the spatial correlation between event occurrences and the regions of local blur, and the temporal correlation between exposure time and the magnitude of global motion blur. To restore highly blurred local areas, we develop a Spatial Implicit Representation Module (SIRM) that models spatial correlation through event encoding sampling to focus on unknown highly blurred areas. To tackle global motion blur, a Temporal Implicit Representation Module (TIRM) is presented to learn temporal correlation via temporal shift operations with long-term aggregation to tackle global motion blur of varying magnitudes. To bridge the real-to-synthetic gap, we finally establish a real-world High-resolution Hybrid Deblur (H2D) dataset with 1920 × 1080 resolution using a new-generation hybrid EVS/CIS sensor, which comprises 1836 group images with naturally spatially aligned and temporally synchronized events at various scales. Qualitative and quantitative experiments demonstrate that the proposed SASNet outperforms eight state-of-the-art methods on both the synthetic GoPro (Nah et al., 2019) and real H2D deblurring datasets, especially in high-speed motion scenarios. The contributions are summarized as follows:

- We propose a Scale-Aware Spatio-temporal deblurring Network (SASNet) to restore unknown highly blurred areas and eliminate global motion blur with varying magnitudes, which is the first time to investigate the arbitrary-scale event-based motion deburring problem to our best knowledge.
- We present spatial and temporal implicit representation modules to model the cross-domain correlation via event encoding sampling and temporal shift operations, which can aggregate correspondence between images and events at arbitrary scales.
- We build a real event-based motion deblurring dataset, High-resolution Hybrid Deblur (H2D), which naturally spatially aligned and temporally synchronized events at various scales using a novel hybrid EVS/CIS sensor. Extensive experiments on both synthetic and real datasets validate the effectiveness of our SASNet.

2. Related Work

2.1. Frame-based Motion Deblurring

Frame-based motion deblurring methods aim to reconstruct clear and sharp images from blurred images. Some early methods (Schmidt et al., 2013; Xu et al., 2014; Chen et al., 2019) adopt classical deconvolution algorithms and introduce additional priors such as total variation to reduce the illness of the reconstruction images. Recently, deep learning-based methods (Kupyn et al., 2018; Cho et al., 2021; Zamir et al., 2021) achieve better performance by directly learning the mapping of blurred images to sharp images. (Chen et al., 2022) presents a nonlinear activation-free network by removing the non-linear activation function to improve the performance. (Zamir et al., 2022) proposes an efficient transformer-based model to capture long-range pixel interactions by designing a multi-head attention network. Despite these advances, these methods are only driven by single-frame data and lack inter-frame motion information, resulting in limited performance.

2.2. Event-based Motion Deblurring

Event-based motion deblurring approaches (Huang et al., 2024; Chen et al., 2024a; Cannici & Scaramuzza, 2024) have attracted widespread attention due to the high temporal resolution of EVS. (Pan et al., 2022) first proposes an Event-based Double Integral (EDI) model to learn the mapping between blurry images, events, and sharp frames, achieving better performance than frame-based methods. (Sun et al., 2022) designs a multi-scale network with a cross-modal attention module, which allows focusing on relevant features of event branches to improve deblurring effects. (Sun et al., 2023) presents a bidirectional recurrent network to fuse information from images and events based on their temporal proximity. (Zhang & Yu, 2022; Zhang et al., 2023b) consider different spatial scales with explicit deformable convolution and handle motion blur via the relativity of

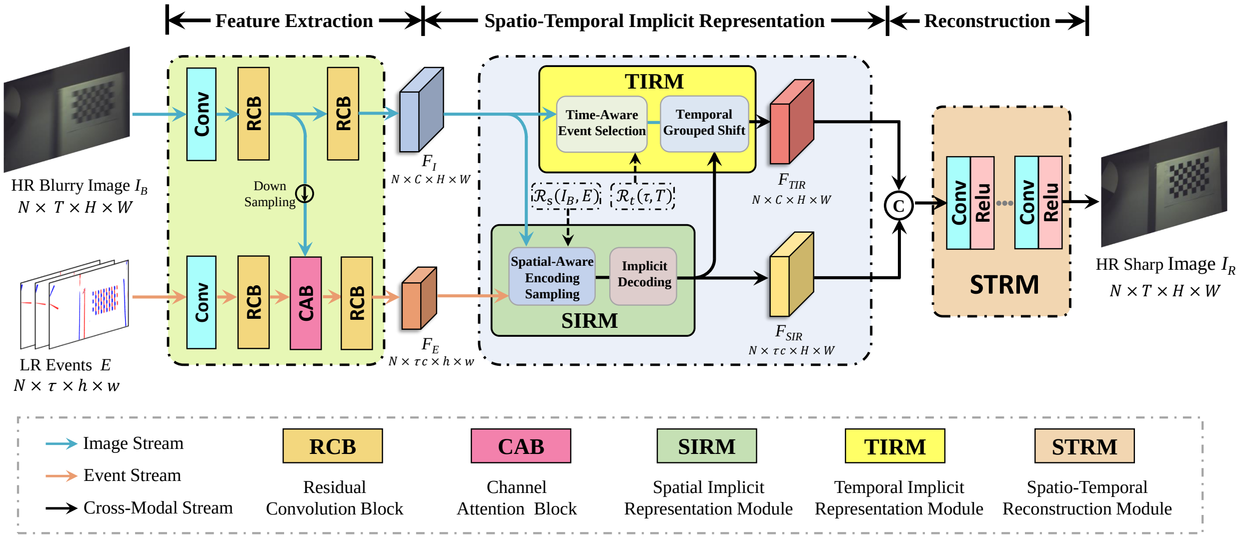

Feature Extraction
Spatio-Temporal Implicit Representation
Reconstruction
HR Blurry Image I_B
$N \times T \times H \times W$
LR Events E
$N \times \tau \times h \times w$
Conv
RCB
RCB
Down Sampling
Conv
RCB
CAB
RCB
F_I
$N \times C \times H \times W$
F_E
$N \times \tau c \times h \times w$
TIRM
Time-Aware Event Selection
Temporal Grouped Shift
$\mathcal{R}_s(I_B, E)$
$\mathcal{R}_t(\tau, T)$
Spatial-Aware Encoding Sampling
Implicit Decoding
SIRM
F_{TIR}
$N \times C \times H \times W$
F_{SIR}
$N \times \tau c \times H \times W$
C
Conv
Relu
Conv
Relu
STRM
HR Sharp Image I_R
$N \times T \times H \times W$
Image Stream
Event Stream
Cross-Modal Stream
RCB
Residual Convolution Block
CAB
Channel Attention Block
SIRM
Spatial Implicit Representation Module
TIRM
Temporal Implicit Representation Module
STRM
Spatio-Temporal Reconstruction Module

Figure 2. Overview of our SASNet, which is composed of three stages: (a) Feature extraction consisting of Residual Convolution Blocks (RCB) and Channel Attention Blocks (CAB). (b) Spatio-Temporal Implicit Representation consists of a Spatial Implicit Representation Module (SIRM) and a Temporal Implicit Representation Module (TIRM). (c) Reconstruction with convolutional layers.

blurriness, which promotes the development of the field. However, these methods (Han et al., 2021; Yu et al., 2023a; Chen et al., 2024a) assume images with the same fixed spatial and temporal scales as events, which cannot deal with real blur with a single model when images and events have different scales. To our knowledge, this paper is the first attempt to study EMD on arbitrary scales.

3. Methodology

3.1. Problem Formulation

We formulate the arbitrary-scale EMD task as a generalized inverse problem, which aims to recover HR sharp images from HR blurry images and LR events at arbitrary spatial and temporal scales. The input HR blurry image $I_B(t)$ is usually formulated as the average of the continuous latent images $I(t)$ at time t within the shutter period T

$$I_B(t) = \frac{1}{T}\int_{t\in(0,T)} I(t)dt \qquad (1)$$

The input LR event streams $E(t, \tau)$ are usually defined by the common EDI model (Pan et al., 2019)

$$E(t,\tau) = \frac{1}{\tau}\int_{t-\frac{\tau}{2}}^{t+\frac{\tau}{2}} \exp(c\int_{t-\Delta t}^{t} p(s)ds)dt \quad \forall t, \tau \in (0, T) \qquad (2)$$

where τ indicates the real exposure duration that falls within a subinterval of shutter period T. Δt denotes the minimum event interval. $p(s) = p \cdot \delta(s - t)$ represents the polarity component of the events. $\delta(\cdot)$ indicates the Dirac function. p denotes the polarity showing the direction of brightness change. Each asynchronous event is emitted when the logarithmic scale brightness change of image $I(t)$ at time t reaches the event threshold c

$$p = \begin{cases} +1, \text{if} \log\left(\frac{I(t)}{I(t-\Delta t)}\right) > c \\ -1, \text{if} \log\left(\frac{I(t)}{I(t-\Delta t)}\right) < -c \end{cases} \qquad (3)$$

To restore the sharp image $I_R(t)$ from $I_B(t)$ and $E(t, \tau)$, EMD task is typically defined as follows

$$I_R(t) = \text{Deblur}(I_B(t), E(t, \tau)) \qquad (4)$$

where $\text{Deblur}(\cdot)$ denotes a deblur network. Due to the uncertainty of spatial resolution and exposure duration in practice, we define the spatial scale as $\mathcal{R}_s(\cdot,\cdot)$ and the temporal scale as $\mathcal{R}_t(\cdot,\cdot)$. $\mathcal{R}_s(I_B(t), E(t, T)) = 2$ means the spatial resolution of the image frame $I_B(t)$ is double times that of events $E(t, T)$. $\mathcal{R}_t(\tau, T) = 0.5$ means the exposure duration of events $E(t, \tau)$ is half times that of the image frame $I_B(t)$. Most previous methods (Sun et al., 2022; 2023) assume that the values of these two scales are fixed at 1. Unlike these methods that focus on fitting Eq. 4, our arbitrary-scale EMD task aims to approximate a more general function

$$\begin{gathered} I_R(t) = \text{Deblur}(I_B(t), E(t, \tau); \mathcal{R}_s, \mathcal{R}_t) \\ \forall t, \tau \in T, \quad \mathcal{R}_s \in [1, 4], \quad \mathcal{R}_t \in (0, 1] \end{gathered} \qquad (5)$$

where $I_R(t)$ is the restored sharp image at time t with exposure time τ and allows flexible input spatial and temporal scales. We set $\mathcal{R}_s$ and $\mathcal{R}_t$ to finite ranges for broadly testing the generalizability of model motion deblurring on the hybrid camera at variable and unknown scales.

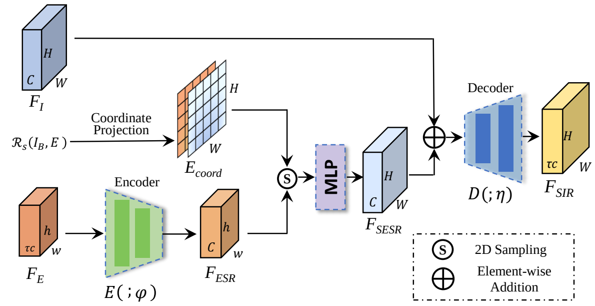


Figure 3. The structure of the proposed Spatial Implicit Representation Module (SIRM).

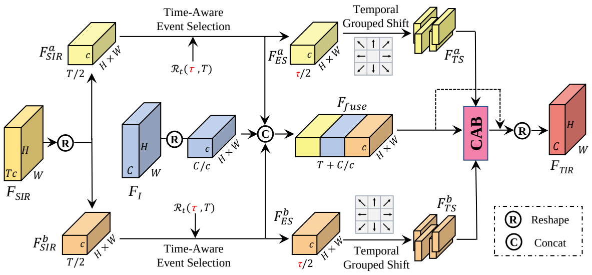


Figure 4. The structure of the proposed Temporal Implicit Representation Module (TIRM).

3.2. Proposed Framework

The overall framework of our SASNet is illustrated in Fig. 2, which efficiently aggregates cross-domain correspondence features of images and events inspired by the spatial-temporal correlation of blurry image formation and event generation. The inputs are an HR blurry image $I_B \in \mathbb{R}^{N\times T\times H\times W}$ and LR events $E \in \mathbb{R}^{N\times\tau\times h\times w}$, where N denotes the batch size. $T = 3$ and $\tau = 16$ indicate the number of channels that collected images and events within those temporal periods. $H \times W$ and $h \times w$ denote spatial sizes of HR and LR, respectively. Note that the input size of spatial and temporal domains can be adjusted according to actual specific requirements. These inputs are first fed to the Residual Convolution Block (RCB) (He et al., 2016) and Residual Attention Block (RAB) (Zhang et al., 2018) to output the image features $F_I \in \mathbb{R}^{N\times C\times H\times W}$ and event features $F_E \in \mathbb{R}^{N\times\tau c\times h\times w}$. Next, the Spatial Implicit Representation Module (SIRM) inputs image and event features to output the spatial implicit representation features $F_{SIR} \in \mathbb{R}^{N\times\tau c\times H\times W}$ according to the arbitrary scale spatial sampling. Then, the image features F_I and F_{SIR} are fed to a Temporal Implicit Representation Module (TIRM) to obtain the temporal implicit representation features $F_{TIR} \in \mathbb{R}^{N\times C\times H\times W}$ by establishing arbitrary-scale temporal correspondences with motion information. Finally, the Spatio-Temporal Reconstruction Module (STRM) with three convolutional layers concatenates F_{SIR} and F_{TIR} with rich spatial-temporal information to reconstruct the HR sharp image $I_R \in \mathbb{R}^{N\times 3\times H\times W}$. In the following, we describe the details of our SIRM and TIRM, respectively.

3.3. Spatial Implicit Representation Module

Due to the influence of object motion speed and depth, the intensity of blurring is unevenly distributed in space. Considering the spatial correlation characteristics of the blurry scene, i.e., the location of the event emitted is highly correlated with the local blurred regions, we propose SIRM to learn arbitrary-scale spatial correspondences to focus on highly blurred areas. As shown in Fig. 3, we adopt an efficient spatial-aware encoding sampling scheme to upsample F_E to desired high-resolution features F_{SESR} and add it to the image features F_I, and then use an implicit decoder parameterized as a Multi-Layer Perceptron (MLP) to convert the added features into continuous feature representation F_{SIR}. Next, we depict the key composition of SIRM.

Spatial-Aware Encoding Sampling. We first encode event features F_E to event spatial features F_{ESR} using an encoder $E(\cdot;\phi)$ with parameters ϕ. Then, a querying coordinate grid $E_{coord} \in \mathbb{R}^{2\times H\times W}$ is generated by coordinate projection according to the input spatial scale R_s. For each coordinate, we select the local features closest to the eight points of this coordinate among the event spatial features F_{EIR} for 2D feature sampling to obtain the sampled spatial features F_{SESR}

$$F_{SESR} = f_{sample}(E(F_E;\phi), E_{coord}) \tag{6}$$

$f_{sample}(\cdot,\cdot)$ denotes the sampling function to dynamically represent features, which can explore the differences between the feature representation abilities of different spatial scales.

Implicit Decoding. The sampled features F_{SESR} and the image features F_I are added together and fed into the implicit decoder $\mathcal{D}(\cdot;\eta)$ with parameters η

$$F_{SIR} = \mathcal{D}(F_{SESR}, F_I;\eta) \tag{7}$$

which decodes the sampled spatial features F_{SESR} to focus on the local texture details by two-layer convolutions for simplicity and efficiency.

3.4. Temporal Implicit Representation Module

The key of the temporal information representation is dependent on the aggregation of inter-frame features, most current works (Xu et al., 2021; Cho et al., 2023) rely on complicated network architectures, e.g., optical flow network (Sun et al., 2018) and transformer self-attention (Zhao et al., 2022; Sabater et al., 2022), resulting in high computational costs. Moreover, their performance is limited when the unknown exposure duration leads to uncertain event temporal frame selection. Therefore, we first select beneficial temporal features from events corresponding to uncertain temporal scales (exposure duration) and then aggregate the selected temporal features and exploit their correlation with motion blur to achieve more efficient global deblurring.

To this end, inspired by the temporal correlation between the exposure time and the degree of motion blur, we propose a simple yet efficient TIRM to implicitly establish temporal correspondences at arbitrary temporal scales, as shown in Fig. 4. TIRM adopts a time-aware grouped shift operation to efficiently aggregate the correspondence feature from bidirectional temporal information. Specifically, we first split the input feature into the pre-shutter neighboring features F^a_{SIR} and the post-shutter neighboring features F^b_{SIR}. The bidirectional event-selected features F^a_{ES} and F^b_{ES} obtained from F^a_{SIR} and F^b_{SIR} through the time-aware event selection with long-term aggregation based on the temporal scale $\mathcal{R}_t$. Then, the F^a_{ES} and F^b_{ES} are fed to temporal grouped shift operations with larger receptive fields to obtain temporal shift features F^a_{TS} and F^b_{TS}. Finally, these shift features and the concatenated fusion features F_{fuse} are fed to a CAB to output the temporal implicit features F_{TIR}.

Time-Aware Event Selection. To effectively maintain the temporal feature within the actual exposure time τ, we leverage time-aware event selection to selectively modulate events based on the input temporal scale $\mathcal{R}_t(\tau, T)$. To achieve long-term bidirectional aggregation, we compress each temporal unit of the shutter neighboring features by applying point-wise convolution, which is formulated as

$$F^a_{ES} = \mathcal{E}(Conv_{1\times1}(F^a_{SIR}), \mathcal{R}_t(\tau, T)) \tag{8}$$

where $\mathcal{E}$ denotes the replication operation along the temporal dimension to select temporal propagating features from neighboring features. It allows adaptively extracting temporal information in arbitrary exposure intervals τ, which alleviates the difficulty of long-term motion estimation.

Temporal Grouped Shift. To tackle global motion blur of varying magnitudes by leveraging event motion information, we present a simple temporal grouped shift operation to implicitly learn correlations between event temporal and motion via large-range spatial shifts. Specifically, the selected features F_{ES} are equally divided into m groups along the temporal dimension, and each unit feature has different shift lengths and directions

$$F^a_{TS,m} = Shift(F^a_{ES,m}, \Delta x_m, \Delta y_m) \tag{9}$$

where $\Delta x_m, \Delta y_m$ represents the shifted pixel in the x and y directions at the m group. In our implementation, we set $m = 4$ and $\Delta x_m, \Delta y_m \in \{-7, -3, 3, 7\}$ to enlarge the receptive field of temporal information aggregation for handling a large range of motion blur. The grouping shift offers multiple potential displacements to match correspondence features at different scales. To seamlessly integrate various shift groups, a CAB is employed with kernel sizes equal to the shift lengths, which achieves a large receptive field and long-term aggregation for global motion deblur.

4. Experiments

4.1. Datasets

For our arbitrary-scale EMD task, two datasets are employed for network training and testing: the synthetic simulated GoPro (Nah et al., 2017) dataset and the real-world H2D dataset built in this paper, which both contain paired LR events, HR blurry images, and HR sharp images.

GoPro Dataset. It consists of 3214 sharp images with resolutions of 1280×720, in which 2103 are used for training and 1111 for testing. We synthesize LR images of various spatial scales from the HR sharp images by bicubic downsampling and then simulate the LR events by using the event camera simulator ICNS (Joubert et al., 2021) at different temporal scales. At the same time, we synthesize the HR blurry images by averaging the adjacent HR sharp images with different frames and different intervals.

H2D Dataset. We collect a High-resolution Hybrid Deblur (H2D) dataset using the novel hybrid EVS/CIS sensor with OV60B (Guo et al., 2023). As shown in Table 2, our H2D dataset consists of 1836 sharp images with resolutions of 1920 $\times 1080$, along with naturally aligned and synchronized LR events from the hybrid CIS with an embedded EVS, where 1233 frames for training and 603 frames for testing. Overall, such a novel bio-inspired hybrid camera enables our H2D to be a competitive dataset with multiple characteristics: (i) high spatial resolution; (ii) natural calibrations in both spatial and temporal domains of images and events; (iii) real-world scenes with abundant diversities in scene category, light change, and movement speed.

4.2. Experimental Settings

The proposed SASNet is implemented by PyTorch and trained on an NVIDIA GeForce RTX 3090 for 100 epochs with 8 batch sizes. The training patch size is set to 256×256 and augmented by horizontal and vertical flipping to enhance its robustness. We use the Adam optimizer (Kingma & Ba, 2014) with an initial learning rate of 10^{-4} that linear decays by 0.5 for every 30 epoch and only employ $\mathcal{L}_1$ loss as the training loss. We compare SASNet with eight state-of-the-art (SOTA) deblurring methods, including frame-based methods (HINet (Chen et al., 2021a), NAFNet (Chen et al., 2022), and Restormer (Zamir et al., 2022)) and event-based methods (EVDI (Zhang & Yu, 2022), UEVD (Kim et al., 2022), EFNet (Sun et al., 2022), and REFID (Sun et al., 2023)). For a fair comparison, all methods are re-trained

Method	Input	GOPRO ($R_s = 4, R_t = 1$)		H2D ($R_s = 2, R_t = 1$)		Complexity	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	#Params	#FLOPs
HINet (Chen et al., 2021a)	Image	25.41	0.7958	32.57	0.9394	88.67M	170.55G
NAFNet (Chen et al., 2022)	Image	27.31	0.8426	32.81	0.9414	16.01M	16.06G
Restormer (Zamir et al., 2022)	Image	28.37	0.8731	33.39	0.9426	26.13M	140.99G
RED-Net (Xu et al., 2021)	Image + Events	27.19	0.8382	33.98	0.9458	9.70M	159.01G
EVDI (Zhang & Yu, 2022)	Image + Events	25.84	0.8069	32.94	0.9432	0.39M	35.54G
UEVD (Kim et al., 2022)	Image + Events	25.69	0.8231	31.98	0.9377	14.23M	101.60G
EFNet (Sun et al., 2022)	Image + Events	28.08	0.8661	34.59	0.9501	8.47M	111.06G
REFID (Sun et al., 2023)	Image + Events	27.51	0.8473	32.61	0.9347	88.96M	208.98G
Ours	Image + Events	28.82	0.8811	35.72	0.9541	1.46M	43.35G

Table 1. Quantitative comparison results of our method and other SOTA methods on the GoPro and H2D datasets. The optimal and suboptimal results are highlighted in red and blue.

	Scenes	Motion	#Seq	#Events (M)	#Images (N)
Indoor	Optical Platform	OM	15	426.5	449
	Pedestrians	OM	5	141.8	152
	Stationary object	CM	15	456.9	451
Outdoor	Traffic	OM	17	459.2	528
	Park Landscape	CM	8	210.8	256
Total			**60**	**1695.2**	**1836**

Table 2. Overview of our real H2D dataset. OM and CM denote the Object Motion and Camera Motion, respectively. #Seqs denotes the number of sequences. #Events indicates the total number of events in the shutter period. #Images represents the number of image frames in the sequence.

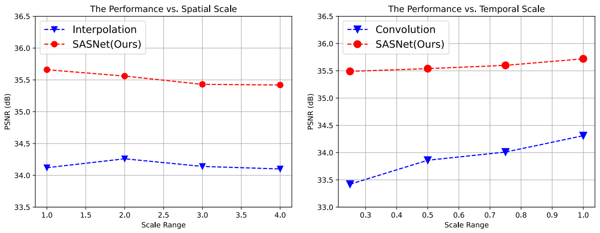
Method	3		6		9	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
EFNet (Sun et al., 2022)	34.59	0.9501	34.26	0.9498	33.86	0.9496
REFID (Sun et al., 2023)	32.61	0.9347	32.19	0.9345	31.81	0.9336
SASNet	35.72	0.9541	35.62	0.9536	35.43	0.9531

Table 3. Quantitative comparison results with other SOTA event-based deblurring methods of different blur frames.

with the same training strategy on our datasets using their official codes. For all experiments, we adopt the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) (Wang et al., 2004) as the evaluation metrics.

4.3. Quantitative Results

As indicated in Table 1, our method with lower model size and computational cost outperforms both frame-based and event-based methods by a large margin on the two datasets. Compared to the event-guided method, our approach consistently outperforms with an average improvement of 1.96 dB on the GoPro dataset and 2.5 dB on the H2D dataset. Compared with UEVD (Zhang & Yu, 2022), SASNet obtains further PSNR gains of 3.13 dB and 3.74 dB for spatial scale factors R_s of 4 and 2, respectively. When compared to the frame-based method, our method with less than one-tenth the number of parameters achieves an average PSNR score improvement of 1.79 dB and 2.79 dB on the GoPro and H2D datasets, respectively. Our SASNet achieves the best performance for all scales and datasets, which indicates our method achieves better results on challenging motion blur frames, especially outstanding significance in the real H2D dataset that has rich textures.



(a) Spatial Scale $\mathcal{R}_s(I_B, E)$

(b) Temporal Scale $\mathcal{R}_t(\tau, T)$

Figure 5. Quantitative results of our SASNet on H2D dataset at different spatial and temporal scales.

Furthermore, we compare the performance of the models under different spatial blur frames in Table 3 to verify their performance for global motion blur. It can be seen that, compared with other algorithms, the performance of our method decreases the least with the increase in blur, especially under the severe blur of frame 9, SASNet only decreases by 0.29 dB (from 35.72 to 35.43), while EFNet decreases by 0.73 dB (from 34.59 to 33.86). It validates the effectiveness of temporal implicit representation with long-term aggregation for different motion magnitudes by establishing arbitrary-scale temporal correspondences with a wide receptive field.

To test the generalization performance of our SASNet on different scales, we also evaluate it on out-of-distribution spatial and temporal scales. Our model is trained on scale ranges of $R_s \in [1, 2]$, $R_t \in (0.5, 1]$, which can generalize to other scale ranges without being re-trained and fine-tuned, making it advantageous for practical applications. It's worth noting that other methods compared in our study often necessitate training multiple models tailored to specific scales, rendering them unsuitable for testing on out-of-distribution

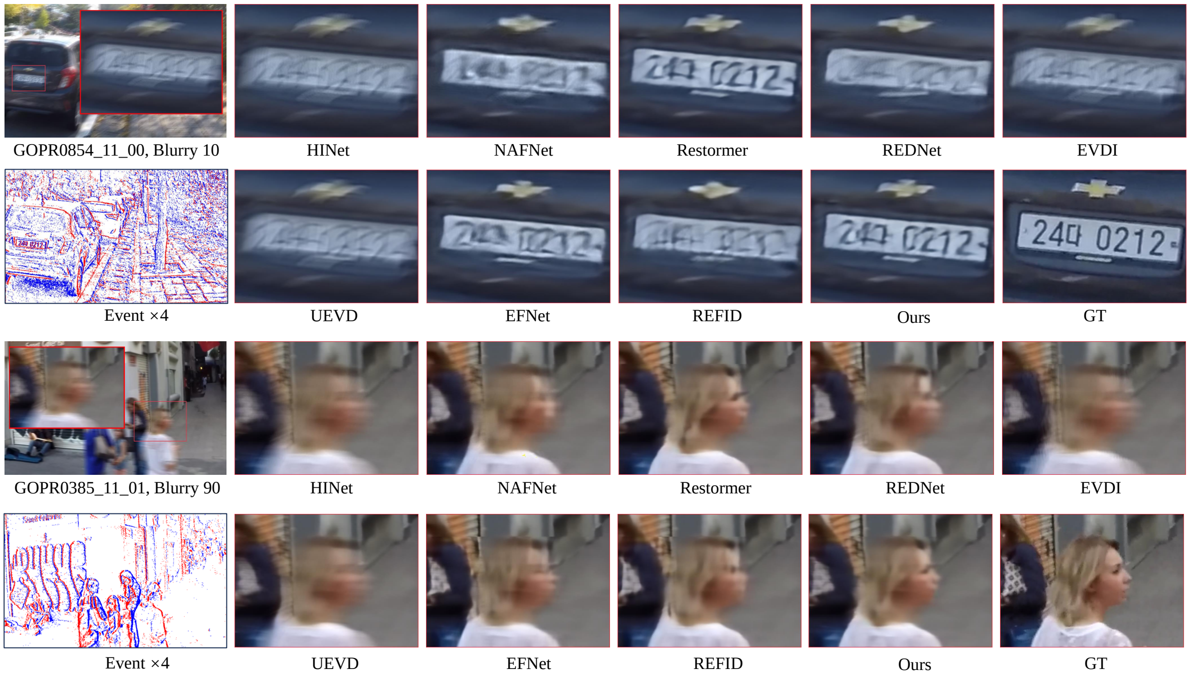


Figure 6. Qualitative comparison results of our method and other SOTA methods on the GoPro dataset. Please zoom in for details.

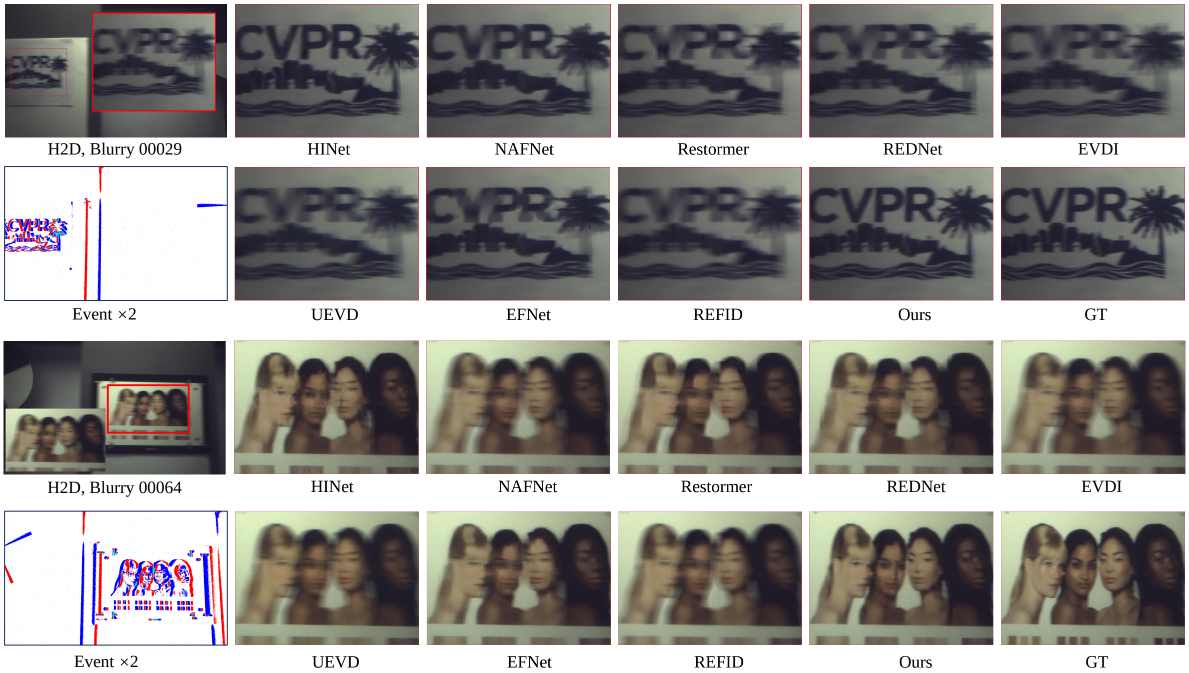


Figure 7. Qualitative comparison results of our method and other SOTA methods on the H2D dataset. Please zoom in for details.

scales. Hence, we selectively employ manual interpolation and varying numbers of convolutional kernels to replace our spatial and temporal implicit operations for fair comparison. As illustrated in Fig. 5, our SASNet remains excellent performance across all scales with a single model. Notably, its performance degradation with increasing scale is minimal, showcasing the exceptional generalization capabilities of our scale-aware spatio-temporal implicit representation for practical applications.

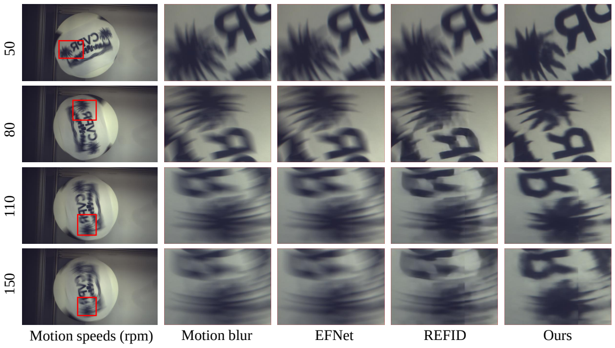


Figure 8. Qualitative comparison results at different rotational speeds. Our method consistently maintains the complete structure of the letter 'R' at high speeds.

Model ID	SIRM	TIRM	PSNR↑	SSIM↑
#1	✗	✗	26.16	0.8418
#2	✓	✗	26.98	0.8569
#3	✗	✓	28.28	0.8701
#4	✓	✓	28.82	0.8811

Table 4. The ablation study of the individual components.

4.4. Qualitative Results

To demonstrate the effectiveness of our method, we present the qualitative comparisons with other methods on the synthesized GoPro dataset in Fig. 6 and the real H2D dataset in Fig. 7. It is evident that HINet and Restormer lose significant detailed textures and exhibit structural distortions in local areas, while EFNet introduces blurred distortions and noise in global areas. In contrast, SASNet achieves more complete structures, fewer blurs, and better detail fidelity, which benefits from the aggregation of spatiotemporal correspondence features by our spatial encoding sampling and temporal correlation learning. Furthermore, we assess the performance of various methods under high-speed motion conditions (i.e., speeds exceeding 50 rpm) in Fig. 8. Our SASNet maintains structural integrity at all speeds, especially excelling in preserving details within highly blurred regions induced by high-speed motion (e.g., 150 rpm).

4.5. Ablation Studies

Individual Components. We perform ablation studies to test the effectiveness of each individual component. All models are trained and evaluated with the same training settings on the GoPro dataset. As shown in Table 4, the result of ID #4 indicates that our spatio-temporal implicit representation achieves an improvement in PSNR by 2.66 dB. SIRM and TIRM achieve average improvements of 0.82 dB and 2.12 dB in PSNR. Due to the different spatial and temporal scales that recover different amounts of information, SIRM and TIRM dynamically adapt to different scales by using scale-aware mechanisms, which achieve continuous feature representation. Next, we conduct additional experiments to verify the abilities of our SIRM and TIRM.

Method	Type	PSNR↑/SSIM↑	#Params	#FLOPs
Interpolation	Explicit	28.36/0.8699	1.461M	43.26G
Transposed Conv	Explicit	28.12/0.8639	1.470M(+0.009)	43.85G(+0.59)
Pixel Shuffle	Explicit	28.43/0.8701	1.610M((+0.149))	43.87G(+0.61)
Learnable Upsample	Implicit	28.61/0.8741	1.688M(+0.227)	43.86G(+0.60)
SIRM (Ours)	Implicit	28.82/0.8811	1.462M(+0.001)	43.35G(+0.09)

Table 5. Comparison of different spatial representation methods on the GoPro dataset. The optimal and suboptimal results are highlighted in red and blue.

Method	Type	PSNR↑/SSIM↑	#Params	#FLOPs
Convolution	Implicit	27.72/0.8631	1.372M	40.39G
Optical Flow	Explicit	27.99/0.8678	1.5387M(0.1667)	51.30G(+10.91)
Deformable Conv	Explicit	28.26/0.8759	1.423M(+0.051)	40.44G(+0.05)
VIT Attention	Implicit	28.84/0.8803	2.514M(+1.142)	77.83G(+37.44)
TIRM (Ours)	Implicit	28.82/0.8811	1.462M(+0.090)	43.35G(+2.96)

Table 6. Comparison of different temporal representation methods on the GoPro dataset. The optimal and suboptimal results are highlighted in red and blue.

In Table 5, using bicubic (Keys, 1981) interpolation as a baseline, we compare our SIRM with three common spatial representation methods: transposed convolution (Dumoulin & Visin, 2016), pixel shuffle (Shi et al., 2016), and learnable upsample (Hu et al., 2019). Explicit methods can only discretely express a single fixed scale and are difficult to generalize to continuous scale intervals. The results indicate that our SIRM achieves the best performance gain (0.46 dB) with the lowest computational complexity (0.09 GFLOPs).

In Table 6, we compare our TIRM with four temporal representation methods, including explicit methods (convolutional networks (Maggioni et al., 2021) and VIT self-attention (Zhao et al., 2022)) and implicit methods (optical flow (Sun et al., 2018) and deformable convolution (Dai et al., 2017)). TIRM requires only 2.96 GFLOPs to achieve performance equivalent to 37.44 GFLOPs of VIT, which is attributed to our straightforward group shift operation. Compared with lightweight deformable convolution, TIRM achieves a performance gain of 0.56 dB (from 28.26 to 28.82), benefiting from the large receptive field provided by the grouped shift operation. The results show that our TIRM achieves an effective balance between computing cost and performance, which validates the excellence of feature shifting for motion deblurring.

Voxel Bins vs. Restoration Performance. Table 7 shows the correlation between the size of the event voxel grid and performance. It can be observed that performance and bin sizes are directly proportional. For a fair comparison, the bin size of 16 in our setting is the same as in EFNet (Sun et al., 2022), yet our SASNet maintains optimal performance.

#Bins	8	12	16	24	32
PSNR↑	25.43	25.51	28.82	28.85	28.97
SSIM↑	0.8750	0.8755	0.8811	0.8820	0.8851

Table 7. Performance with different event voxel grid bin sizes.

5. Conclusions

In this paper, we investigate the challenge of event-based motion deblurring on arbitrary spatial and temporal scales, which is practical yet has never been investigated so far. To conquer this challenge, we collect an H2D dataset by using a hybrid EVS/CIS camera, comprising images with naturally spatially aligned and temporally synchronized events at various scales, which is essential to our arbitrary-scale EMD task. Then, we implement a novel Scale-Aware Spatio-temporal Network (SASNet) to flexibly restore blurred images with event streams at arbitrary scales, which implicitly aggregate both spatial and temporal correspondence features of images and events to focus on unknown severe local blur and remove global motion blur of varying magnitudes. To restore highly blurred local areas, we propose a Spatial Implicit Representation Module (SIRM) to aggregate spatial correlation at any resolution through event encoding sampling. To tackle global motion blur, a Temporal Implicit Representation Module (TIRM) is presented to learn temporal correlation via temporal shift operations with long-term aggregation. As validated by comprehensive experiments on synthetic and real datasets, we demonstrate the superior performance of our SASNet in dealing with arbitrary-scale EMD problems.

Despite the fact that our SASNet achieves optimal quantitative and qualitative results across various datasets, its application to real-world situations encounters challenges beyond solely motion blur problems. Consequently, future research must delve into enhancing the algorithm's robustness by incorporating a broader range of degraded real-world data into both the training and evaluation processes.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 62272134 and 62072141), and the National Key RD Program of China under Grant (No. 2018AAA0102801).

Impact Statement

This paper presents a scale-aware spatio-temporal implicit representation algorithm for the event-based motion deblurring task. Our work makes significant contributions to image deblurring, event representation, and computational imaging. This paper can deal with event-based deblurring in practice at arbitrary scales. Given the ever-increasing size of datasets and the advent of the era of large models, our research offers a practical and viable technological solution for enhancing computational resource utilization. While there may be many other potential impacts, we believe that our approach does not entail any negative ethical or moral implications.

References

Cannici, M. and Scaramuzza, D. Mitigating motion blur in neural radiance fields with events and frames. *arXiv preprint arXiv:2403.19780*, 2024.

Chen, K., Chen, S., Zhang, J., Zhang, B., Zheng, Y., Huang, T., and Yu, Z. Spikereveal: Unlocking temporal sequences from real blurry inputs with spike streams. *arXiv preprint arXiv:2403.09486*, 2024a.

Chen, L., Fang, F., Wang, T., and Zhang, G. Blind image deblurring with local maximum gradient prior. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1742–1750, 2019.

Chen, L., Lu, X., Zhang, J., Chu, X., and Chen, C. Hinet: Half instance normalization network for image restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 182–192, 2021a.

Chen, L., Chu, X., Zhang, X., and Sun, J. Simple baselines for image restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 17–33, 2022.

Chen, S., Zhang, J., Zheng, Y., Huang, T., and Yu, Z. Enhancing motion deblurring in high-speed scenes with spike streams. *Advances in Neural Information Processing Systems*, 36, 2024b.

Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., and Liu, Z. Dynamic convolution: Attention over convolution kernels. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 11030–11039, 2020.

Chen, Y., Liu, S., and Wang, X. Learning continuous image representation with local implicit image function. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8628–8638, 2021b.

Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X., and Yu, F. Dual aggregation transformer for image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12312–12321, 2023.

Cho, H., Jeong, Y., Kim, T., and Yoon, K.-J. Non-coaxial event-guided motion deblurring with spatial alignment. In *Int. Conf. Comput. Vis.*, pp. 12492–12503, 2023.

Cho, S.-J., Ji, S.-W., Hong, J.-P., Jung, S.-W., and Ko, S.-J. Rethinking coarse-to-fine approach in single image deblurring. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4641–4650, 2021.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y. Deformable convolutional networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., and Brox, T. Flownet: Learning optical flow with convolutional networks. In *Int. Conf. Comput. Vis.*, pp. 2758–2766, 2015.

Dumoulin, V. and Visin, F. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

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

Gao, Y., Li, S., Li, Y., Guo, Y., and Dai, Q. Superfast: 200× video frame interpolation via event camera. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

Gu, J. and Dong, C. Interpreting super-resolution networks with local attribution maps. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9199–9208, 2021.

Guo, M., Chen, S., Gao, Z., Yang, W., Bartkovjak, P., Qin, Q., Hu, X., Zhou, D., Uchiyama, M., Fukuoka, S., et al. A 3-wafer-stacked hybrid 15mpixel cis+ 1 mpixel evs with 4.6 gevent/s readout, in-pixel tdc and on-chip isp and esp function. In *Int. Solid-State Circuits Conf.*, pp. 90–92. IEEE, 2023.

Han, J., Yang, Y., Zhou, C., Xu, C., and Shi, B. Evintsr-net: Event guided multiple latent frames reconstruction and super-resolution. In *Int. Conf. Comput. Vis.*, pp. 4882–4891, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 770–778, 2016.

Hidalgo-Carrió, J., Gallego, G., and Scaramuzza, D. Event-aided direct sparse odometry. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5781–5790, 2022.

Hu, X., Mu, H., Zhang, X., Wang, Z., Tan, T., and Sun, J. Meta-sr: A magnification-arbitrary network for super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1575–1584, 2019.

Huang, Z., Liang, Q., Yu, Y., Qin, C., Zheng, X., Huang, K., Zhou, Z., and Yang, W. Bilateral event mining and complementary for event stream super-resolution. *arXiv preprint arXiv:2405.10037*, 2024.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Joubert, D., Marcireau, A., Ralph, N., Jolley, A., van Schaik, A., and Cohen, G. Event camera simulator improvements via characterized parameters. *Adv. Neural Inform. Process. Syst.*, 15:702765, 2021.

Keys, R. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160, 1981.

Kim, T., Lee, J., Wang, L., and Yoon, K.-J. Event-guided deblurring of unknown exposure time videos. In *European Conference on Computer Vision*, pp. 519–538, 2022.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., and Matas, J. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8183–8192, 2018.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

Maggioni, M., Huang, Y., Li, C., Xiao, S., Fu, Z., and Song, F. Efficient multi-stage video denoising with recurrent spatio-temporal fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3466–3475, 2021.

Nah, S., Hyun Kim, T., and Mu Lee, K. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3883–3891, 2017.

Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., and Mu Lee, K. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pp. 1974–1984, 2019.

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., and Dai, Y. Bringing a blurry frame alive at high frame-rate with an event camera. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 6820–6829, 2019.

Pan, L., Hartley, R., Scheerlinck, C., Liu, M., Yu, X., and Dai, Y. High frame rate video reconstruction based on an event camera. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(5):2519–2533, 2022.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

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

Sabater, A., Montesano, L., and Murillo, A. C. Event transformer. a sparse-aware solution for efficient event data processing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2677–2686, 2022.

Schmidt, U., Rother, C., Nowozin, S., Jancsary, J., and Roth, S. Discriminative non-blind deblurring. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 604–611, 2013.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1874–1883, 2016.

Sun, D., Yang, X., Liu, M.-Y., and Kautz, J. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8934–8943, 2018.

Sun, L., Sakaridis, C., Liang, J., Jiang, Q., Yang, K., Sun, P., Ye, Y., Wang, K., and Gool, L. V. Event-based fusion for motion deblurring with cross-modal attention. In *European Conference on Computer Vision*, pp. 412–428, 2022.

Sun, L., Sakaridis, C., Liang, J., Sun, P., Cao, J., Zhang, K., Jiang, Q., Wang, K., and Van Gool, L. Event-based frame interpolation with ad-hoc deblurring. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 18043–18052, 2023.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004.

Xu, F., Yu, L., Wang, B., Yang, W., Xia, G.-S., Jia, X., Qiao, Z., and Liu, J. Motion deblurring with real events. In *Int. Conf. Comput. Vis.*, pp. 2583–2592, 2021.

Xu, L., Tao, X., and Jia, J. Inverse kernels for fast spatial deconvolution. In *Eur. Conf. Comput. Vis.*, pp. 33–48, 2014.

Yang, D. and Yamac, M. Motion aware double attention network for dynamic scene deblurring. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1113–1123, 2022.

Yu, H., Li, H., Yang, W., Yu, L., and Xia, G.-S. Detecting line segments in motion-blurred images with events. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023a.

Yu, L., Wang, B., Zhang, X., Zhang, H., Yang, W., Liu, J., and Xia, G.-S. Learning to super-resolve blurry images with events. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45 (8):10027–10043, 2023b.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H., and Shao, L. Multi-stage progressive image restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., and Yang, M.-H. Restormer: Efficient transformer for high-resolution image restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5728–5739, 2022.

Zhang, C., Zhang, X., Lin, M., Li, C., He, C., Yang, W., Xia, G.-S., and Yu, L. Crosszoom: Simultaneously motion deblurring and event super-resolving. *arXiv preprint arXiv:2309.16949*, 2023a.

Zhang, X. and Yu, L. Unifying motion deblurring and frame interpolation with events. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 17765–17774, 2022.

Zhang, X., Yu, L., Yang, W., Liu, J., and Xia, G.-S. Generalizing event-based motion deblurring in real-world scenarios. In *Int. Conf. Comput. Vis.*, pp. 10734–10744, 2023b.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., and Fu, Y. Image super-resolution using very deep residual channel attention networks. In *Eur. Conf. Comput. Vis.*, pp. 286–301, 2018.

Zhang, Z., Yezzi, A., and Gallego, G. Formulating event-based image reconstruction as a linear inverse problem with deep regularization using optical flow. *IEEE Trans. Pattern Anal. Mach. Intell.*, (01):1–18, 2022.

Zhao, J., Zhang, S., and Huang, T. Transformer-based domain adaptation for event data classification. In *Int. Conf. Acoust., Speech, Signal Process.*, pp. 4673–4677, 2022.

A. Appendix Outline

In this supplementary material, we provide detailed descriptions of our arbitrary-scale event-based motion deblurring solution as follows:

1) Sec. B provides the network parameters and experimental settings of our SASNet.
2) Sec. C describes more details about the newly built hybrid event-based motion deblurring dataset (H2D).
3) Sec. D presents additional ablation studies to further verify the effectiveness of our key SIRM and TIRM.
4) Sec. E reports more visualization comparison results at different scales to verify model generalization.

The source code and collected dataset will also be released upon publication.

B. Architecture and Experimental Settings

Architecture. As shown in Fig. 2, our Scale-Aware Spatio-temporal Network (SASNet) is composed of several simple convolutional blocks and modules. Specifically, the Residual Convolution Blocks (RCB) (He et al., 2016) and Channel Attention Blocks (CAB) (Zhang et al., 2018) are commonly used convolutional blocks for feature extraction. RCB is a four-layer convolutional with ReLU activation (Nair & Hinton, 2010) and hidden dimensions of 32. CAB is a two-layer convolutional and an additional channel attention layer with the hidden dimensions set to 64. The architecture of our Spatial Implicit Representation Module (SIRM) is shown in Fig. 3, where the encoder $E(\cdot;\phi)$ and the decoder $\mathcal{D}(\cdot;\eta)$ are both composed of a two-layer MLP with ReLU activation and hidden dimensions of 32. The architecture of our Temporal Implicit Representation Module (TIRM) is shown in Fig. 4, where time-aware event selection only adopts a one-layer 1 $\times$ 1 convolution and CAB has eight-layer convolutional with channel attention layers. The application of the SIRM and TIRM with low computational cost can extend the spatial and temporal scales beyond the distribution and achieve better performance due to their implicit continuous feature expression.

Experimental Settings. In this section, we present the details of the training process for our arbitrary-scale Event-based Motion Deblurring (EMD) method. First, we denote $N = 8$ as the batch size and sample random spatial scales $\mathcal{R}_s$ in uniform distribution $\mathcal{U}(1, 4)$ and random temporal scales $\mathcal{R}_t$ in uniform distribution $\mathcal{U}(0, 1)$. Then, we crop N patches with the size $\{256 \times 256\}$ from training HR sharp and blurred images and the size of $\{256/\mathcal{R}_s \times 256/\mathcal{R}_s\}$ from events within $\mathcal{R}_t$ time duration as the training pairs. We input images and events into the networks in pairs to calculate the $\mathcal{L}1$ loss function. Finally, to accelerate convergence, we adopt the SWA (Izmailov et al., 2018) strategy to update model parameters with an update interval of 10, making the training more robust and achieving wider generalization at different spatial and temporal scales.

C. H2D Dataset

To the best of our knowledge, there are no publicly released datasets available for our arbitrary-scale EMD task with data at different temporal and spatial scales, which motivates us to build a new hybrid dataset containing paired clear-blurred images and aligned events in both spatial and temporal domains for practical requirements.

Hybrid EVS/CIS Sensor. Existing commonly used dual-sensor acquisition systems (Sun et al., 2022; Zhang et al., 2023b; Cho et al., 2023) have several shortcomings, including parallax errors in camera collocation, the complexities of spatial and temporal synchronization of CMOS image sensor (CIS) and event-based vision sensor (EVS), and increased cost. To reduce complexity and narrow the gap between real and synthetic, we adopt a novel hybrid EVS/CIS sensor OV60B (Guo et al., 2023), consisting of a hybrid 1920 × 1080 CIS with an embedded 960 × 540 EVS operating at 120 frames per second. Within a 4 × 4 cluster of CIS pixels, one color channel is replaced to provide photocurrent to the EVS pixels, which overcomes the above shortcomings and can deliver the high spatial resolution event requirements without significantly sacrificing CIS performance. Since both CIS and EVS are integrated on the same wafer, both the LR events and HR images in the spatial and temporal domains are naturally aligned, i.e., share the same spatial field of view and timestamps, which avoids manual alignment and synchronization of data at different spatial and temporal scales. Note that calibration is crucial for our arbitrary-scale EMD task.

Dataset	Color	Camera	Image Resolution	Event Resolution	Type of Scenes	SA	TS	HR
BS-ERGB	RGB	FLIR + Prophesee Gen4	970 × 625	970× 625	Low Speed	✗	✗	✓
THU-HSEVI	Gray	EoSens + DAVIS346	340× 260	340× 260	High Speed	✗	✗	✗
DAVIS 240C Dataset	Gray	DAVIS246	240× 180	240× 180	Low Speed	✓	✓	✗
HQF	Gray	DAVIS240C	346× 260	346× 260	Low Speed	✓	✓	✗
Ours (H2D)	RGB	OV60B	1920× 1080	960× 540	High Speed	✓	✓	✓

Table 8. The Comparison of our H2D Dataset with other event-based deblurring datasets. "SA" denotes Spatial Alignment, "TS" denotes Temporal Synchronization, "HR" indicates High-Resolution.

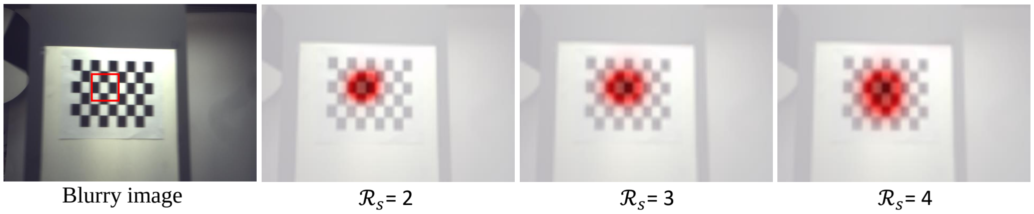

Blurry image $\mathcal{R}_s= 2$ $\mathcal{R}_s= 3$ $\mathcal{R}_s= 4$

Figure 9. Visualization of dynamic receptive fields for our SIRM at different spatial scales.

Data Collecting. Our H2D dataset contains a total of 60 sequences, including spatio-temporal aligned low-resolution (LR) event streams and high-resolution (HR) image sequences, with abundant diversity in terms of scenes (i.e., indoor and outdoor) and motion types (i.e., camera and object motion) at different light changes and motion speeds. The HR sharp images are captured with a steady tripod-mounted camera at stable illuminations and smooth movements. The HR motion-blurred images are synthesized by averaging adjacent three sharp frames, resulting in 1836 pairs of blurred and sharp images. Simultaneously, LR event streams with different exposure times and spatial resolutions are read out from the hybrid sensor through parameter settings. To further improve model generalization, we utilize bicubic resizing in PyTorch (Paszke et al., 2019) by taking the center point as the basement to perform arbitrary scale down-sampling for training and testing at arbitrary scales. We randomly select 40 sequences with 1233 pairs as the training set and the remaining 20 with 603 pairs as the testing set.

In Table 8, we show the comparison of our H2D with other event-based deblurring datasets. The high resolution of the H2D dataset is crucial for the evaluation and benchmarking of our arbitrary-scale event-based deblurring task under more challenging high-speed and complex conditions, leading to more robust solutions in real-world applications.

D. Quantitative Ablation Study

To further verify the effectiveness of our SIRM, we decouple SIRM to obtain a Spatial-Aware Encoding Sampling (SAES) and an Implicit Decoding for detailed analysis in Table 9. With 'SAES', our method adaptively adjusts the receptive field range of the fusion network from events to images according to the input spatial scale $\mathcal{R}_s$, which achieves a 0.27 dB improvement in terms of PSNR on the GoPro (Nah et al., 2019) dataset. Different from the traditional explicit upsampling method (Shi et al., 2016; Dumoulin & Visin, 2016), our implicit decoding maps the coordinate values of discrete pixels in two-dimensional space to pixel values, and adopts the MLP function (Chen et al., 2021b) to represent the image to arbitrary spatial scales continuously. Furthermore, we adopt LAM (Gu & Dong, 2021) to visualize the receptive field of our SIRM at different spatial scales, as shown in Fig. 9. The results demonstrate that the dynamic receptive field of local regions with complex texture structures has coincided with the spatial scale, i.e., a higher resolution scale corresponds to a larger receptive field range. Thus, our SIRM models the spatial correlation through implicit representation on different spatial scales.

Moreover, to verify the ability of our TIRM, we decouple TIRM to obtain a Time-Aware Event Selection (TAES) and a Temporal Grouped Shift (TGS) for detailed analysis in Table 10. Specifically, we conduct a series of comparative experiments in terms of performance and efficiency to further elaborate on the advantages and functions of the TAES and TGS. According to the experiment results in Table 10, we can find that TAES and TGS improve performance by 0.44 dB and 0.81 dB respectively. To select beneficial events and reduce subsequent computational costs, TAES uses a layer of point-wise convolution to avoid the loss of spatial information with a negligible computational complexity of 0.004 GFLOPs.

SIRM	PSNR↑	SSIM↑
w/o SAES	28.45	0.8712
w SAES	28.82	0.8811

Table 9. Ablation experiment results of SIRM.

TIRM		PSNR↑	SSIM↑	#Params	#Flops
TAES	TGS				
✗	✗	27.72	0.8631	1.372M	40.39G
✓	✗	28.16	0.8693	1.381M	40.39G
✗	✓	28.53	0.8736	1.453M	43.35G
✓	✓	28.82	0.8811	1.462M	43.35G

Table 10. Ablation experiment results of our TIRM.

Reconstruction Model	PSNR	SSIM	#Params (M)	#Flops (G)	#Times (ms)
MPRNet (Zamir et al., 2021)	28.78	0.8726	2.15	84.27	70.99
SwinT (Liang et al., 2021)	28.86	0.8715	1.77	94.76	113.35
DAT (Chen et al., 2023)	28.89	0.8813	2.38	149.27	116.44
Ours	28.82	0.8811	1.46	43.35	58.03

Table 11. Comparison Experiments of Reconstruction Modules.

To learn the correlation between event temporal sequences and motion blur, TGS performs multiple groups of large-range feature shifts to implicitly implement temporal feature representation and fusion. It can be attributed to the proposed event selection and temporal shift, which can implicitly learn the bidirectional temporal correlation before and after, and achieve long-range feature aggregation through a large receptive field.

To validate the effectiveness of the spatial-temporal reconstruction module, we conduct additional experiments to compare it with three other reconstruction modules. As shown in Table 11, although DAT achieves the best performance in terms of PSNR and SSIM, its computational cost is about three times that of ours (149.27 GFLOPs vs. 43.35 GFLOPs). Overall, the proposed spatial-temporal reconstruction module achieves comparable performance while significantly reducing the number of parameters and computations.

E. Additional Qualitatuve Experiments

In Figs. 11 - 14, we present more visual comparison results of our SASNet and other SOTA methods on GoPro and H2D datasets at different scales. From Figure. 11 to Figure. 13, we show the results of examples from our H2D dataset. To clearly reflect the difference in the visual results of these algorithms, we chose different decimal scales of ×1.5, ×2.5, and ×3.5 for visualization. As shown in Figs. 11 and 12, our algorithm can effectively remove global blur and accurately reconstruct more realistic spatial structures and more texture details. Other existing algorithms suffer from loss of detail and distortion of the structure. In Figure. 13, we see that SASNet outperforms in both fidelity and quality and perceives the surrounding repetitive structure to predict the accurate target area. It is noteworthy that the performance of REFID (Sun et al., 2023) is less promising in the category of architecture texture. In comparison with REDNet (Xu et al., 2021) and EFNet (Sun et al., 2022) on spatial structure texture, SASNet produces realistic outputs that are highly similar to the ground truths. This proves that the scale-aware receptive field is very important for arbitrary-scale reconstruction tasks. In general, guided by the scale-aware implicit representation mechanism, our SASNet better utilizes motion blur at different scales for performance generalization to eliminate both local and global motion blur and achieve high fidelity.

As shown in Table 12, our method with event data achieves PSNR improvements by 0.66 dB and 1.29 dB on GoPro and our H2D datasets, respectively, which indicates events are very useful in assisting RGB image deblurring. Since our H2D dataset includes complex scenes with different motion speeds and different lighting conditions, we further conduct a visual

	GOPRO		H2D	
Event	PSNR↑	SSIM↑	PSNR↑	SSIM↑
w/	28.82	0.8811	35.72	0.9541
w/o	28.16	0.8128	34.41	0.9461

Table 12. The effectiveness of events on image deblurring.

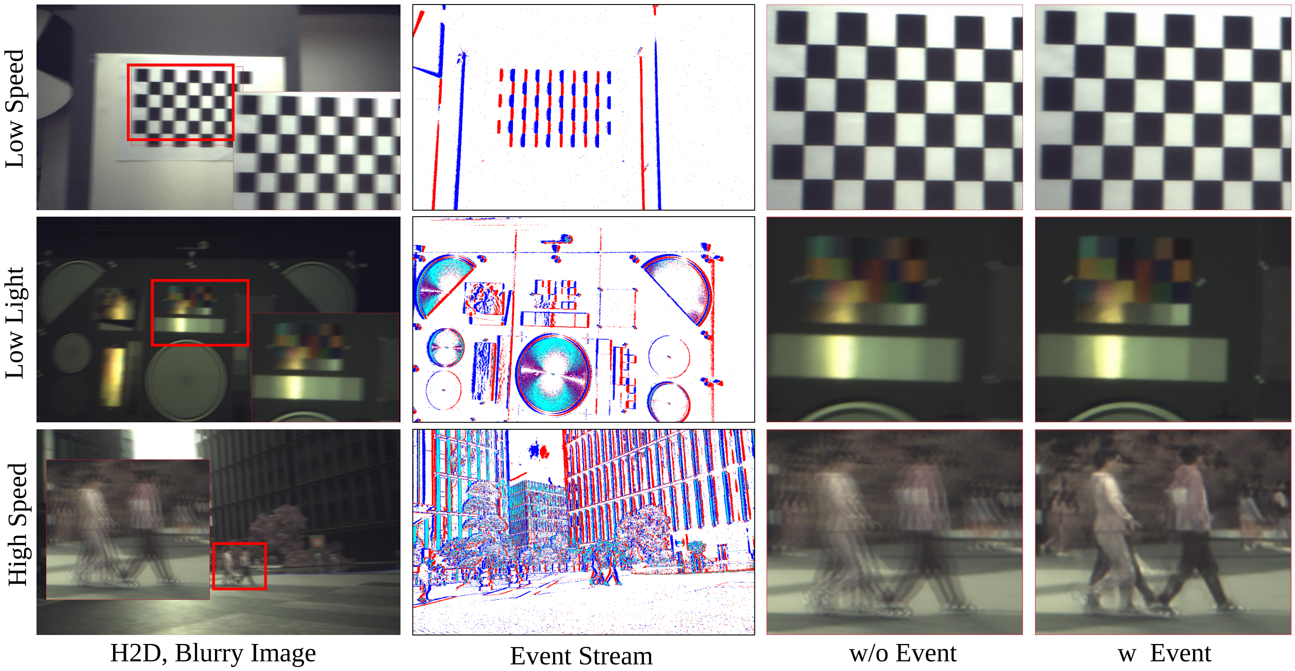


Figure 10. The performance comparison between our method with and without event data.

comparative analysis.

As shown in Figure 10, our method without event data can also achieve good restoration results in slow-speed motion scenes. However, our method with event data causes significant performance improvement in high-speed motion and low-light scenarios. The reason is that the event stream captures brightness changes in dynamic scenes at a very high temporal resolution. When image blur is caused by high-speed motion, the temporal characteristics of events can accurately track the trajectory of the moving objects, which helps resolve global and large-scale motion blur. Meanwhile, due to the high dynamic range of event cameras, even under low lighting conditions, the spatial features of events can still capture high-contrast edges of objects, which helps compensate for fine textures affected by local motion blur. Therefore, event data enhances the deblurring performance of algorithms in many complex scenes.

F. Future works

While SASNet shows better capability compared to existing works, the current degradation only considers motion blur, which mismatches the more complex degradation requirements in real-world reconstruction (such as noise, atmospheric scattering, *etc*). Further restoration generalization ability of our SASNet in real-world applications will be explored in future works. And, our H2D dataset and framework can be flexibly extended to other low-level event-based vision tasks, such as super-resolution, low-light enhancement, and video frame interpolation, to achieve more practical applications.

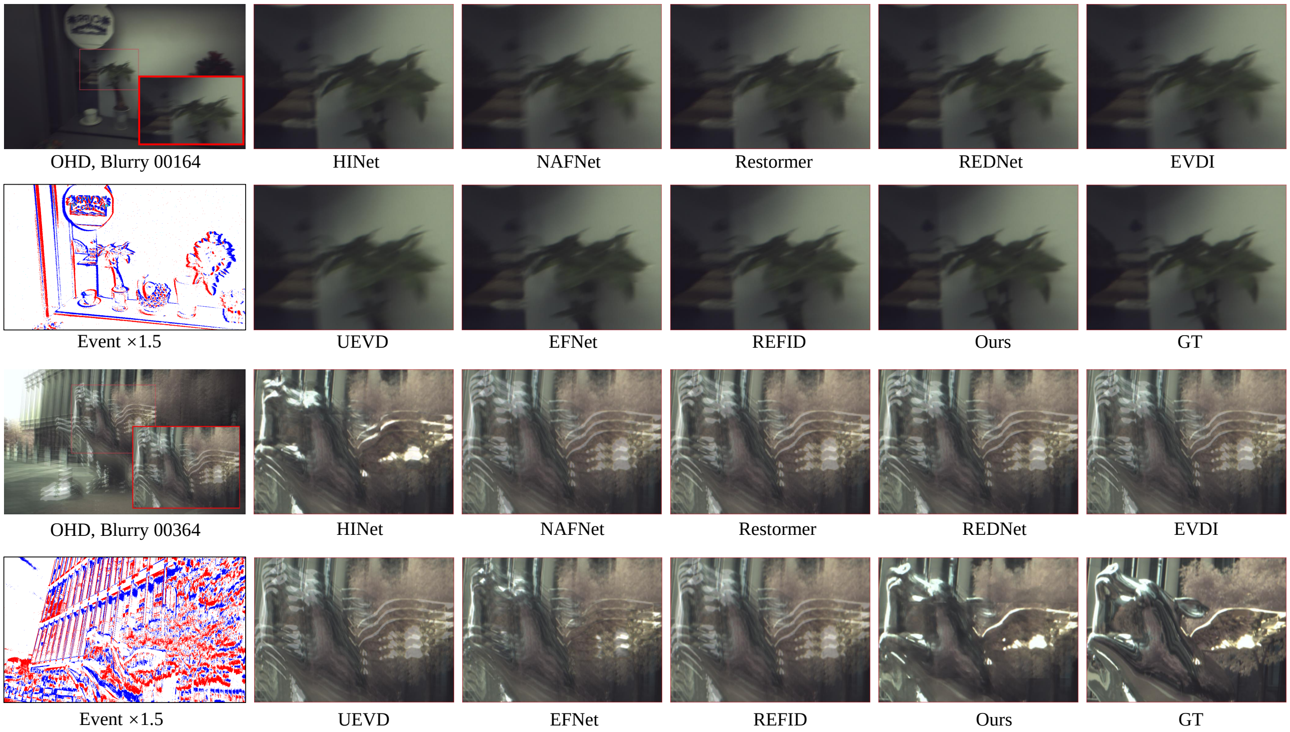


Figure 11. Qualitative comparison results of our method and other SOTA methods on the H2D images synthesized through interval frames. Please zoom in for details.

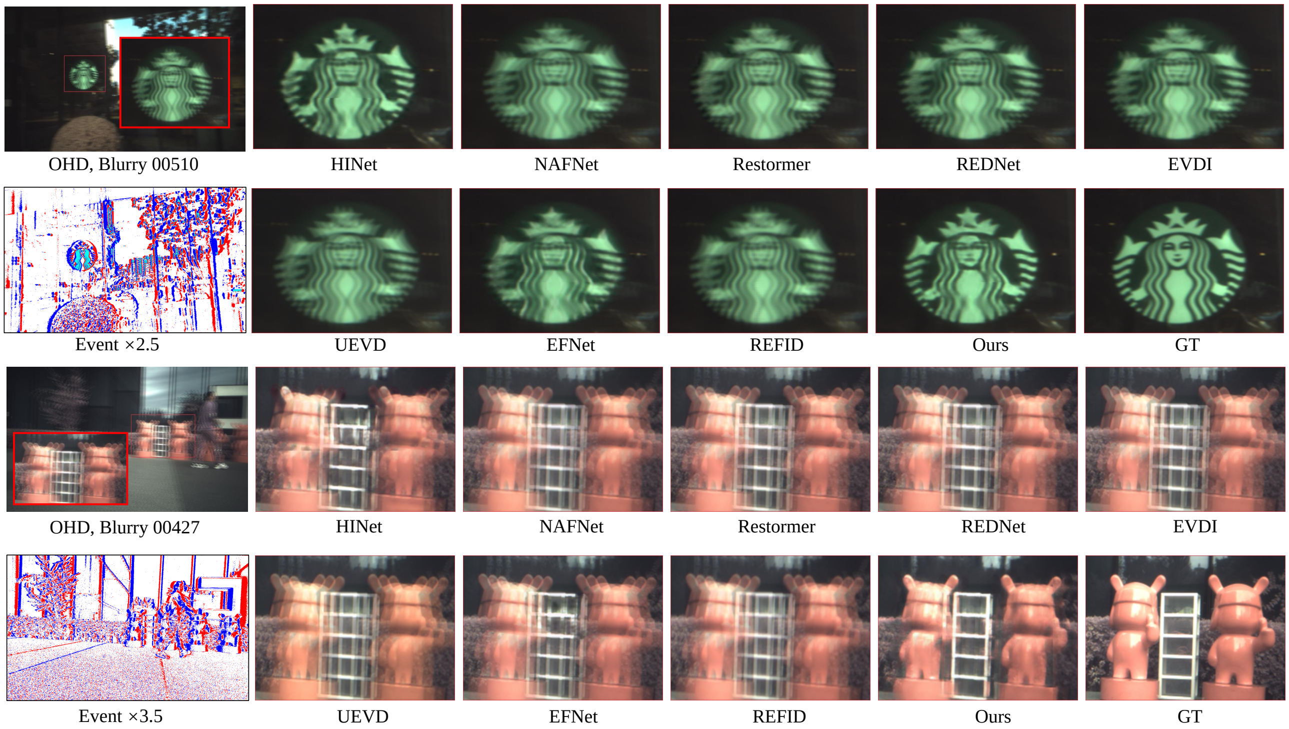


Figure 12. Qualitative comparison results of our method and other SOTA methods on the H2D images synthesized through interval frames. Please zoom in for details.

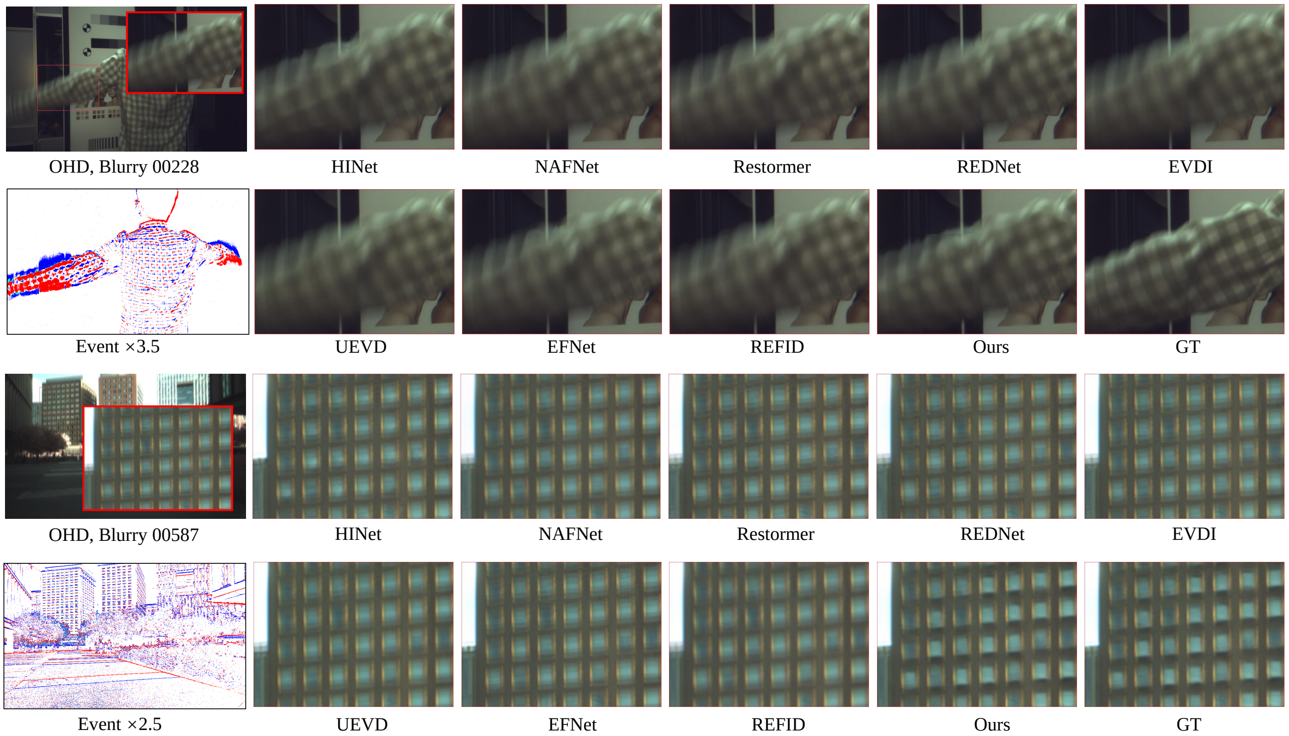


Figure 13. Qualitative comparison results of our method and other SOTA methods on the H2D images synthesized through continuous frames. Please zoom in for details.

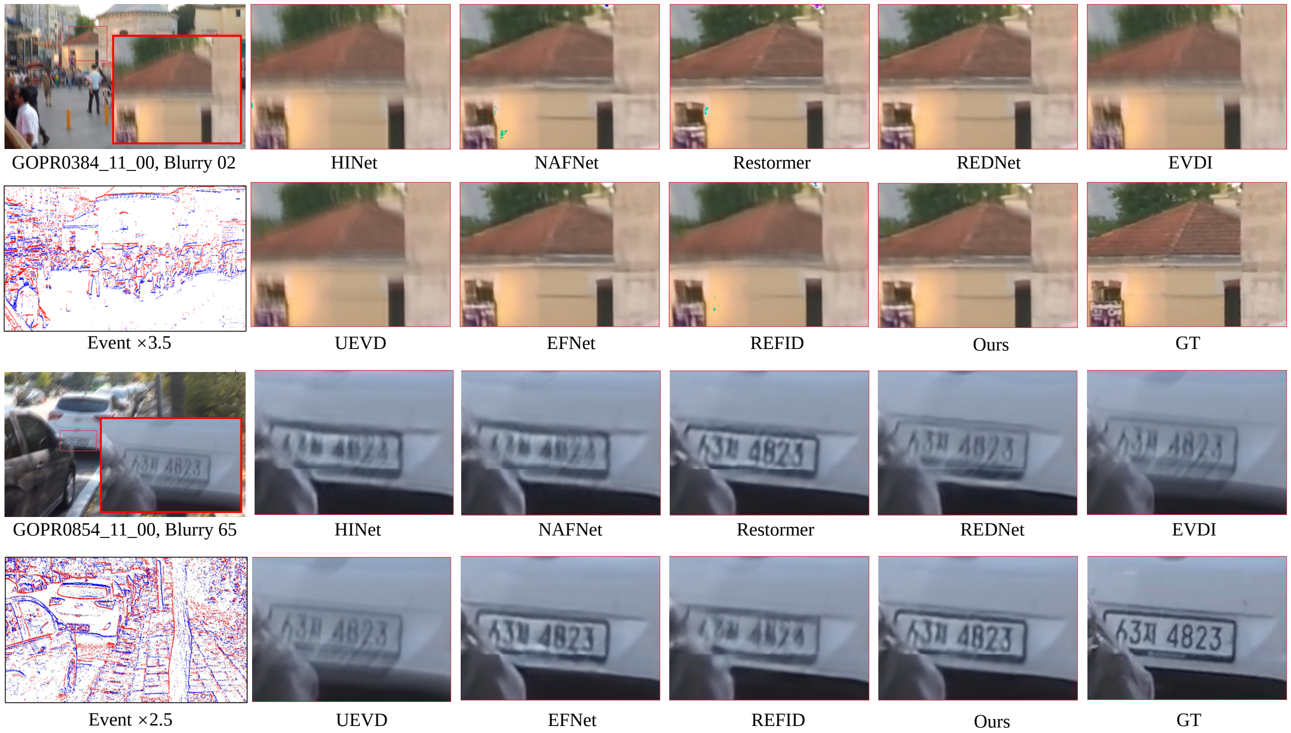


Figure 14. Qualitative comparison results of our method and other SOTA methods on the GOPRO images synthesized through continuous frames. Please zoom in for details.