ClearSight: Human Vision-Inspired Solutions for Event-Based Motion Deblurring

Xiaopeng Lin*, Yulong Huang*, Hongwei, Ren, Zunchang Liu, Yue Zhou, Haotian Fu, Bojun Cheng✉
The Hong Kong University of Science and Technology (Guangzhou)
{xlin746, yhuang496, hren066, zliu361, yzhou883, hf373}@connect.hkust-gz.edu.cn
bocheng@hkust-gz.edu.cn

Abstract

Motion deblurring addresses the challenge of image blur caused by camera or scene movement. Event cameras provide motion information that is encoded in the asynchronous event streams. To efficiently leverage the temporal information of event streams, we employ Spiking Neural Networks (SNNs) for motion feature extraction and Artificial Neural Networks (ANNs) for color information processing. Due to the non-uniform distribution and inherent redundancy of event data, existing cross-modal feature fusion methods exhibit certain limitations. Inspired by the visual attention mechanism in the human visual system, this study introduces a bioinspired dual-drive hybrid network (BDHNet). Specifically, the Neuron Configurator Module (NCM) is designed to dynamically adjusts neuron configurations based on cross-modal features, thereby focusing the spikes in blurry regions and adapting to varying blurry scenarios dynamically. Additionally, the Region of Blurry Attention Module (RBAM) is introduced to generate a blurry mask in an unsupervised manner, effectively extracting motion clues from the event features and guiding more accurate cross-modal feature fusion. Extensive subjective and objective evaluations demonstrate that our method outperforms current state-of-the-art methods on both synthetic and real-world datasets.

1. Introduction

Motion blurring primarily occurs due to the movement of either the camera or the moving objects during the sensor's exposure period [19, 20]. Deblurring is a critical task focused on recovering a sharp image with clear details from the motion-blurred counterpart. Several image-based deblurring approaches have been developed to compensate for the blur characteristics with enhanced performance, including traditional approaches[12, 2] and learning-based approaches [9, 31, 8]. However, deblurring methods that rely solely on conventional frame-based cameras frequently face performance constraints due to the absence of essential motion information. These limitations are particularly pronounced under adverse lighting conditions and during the capture of rapidly moving objects.

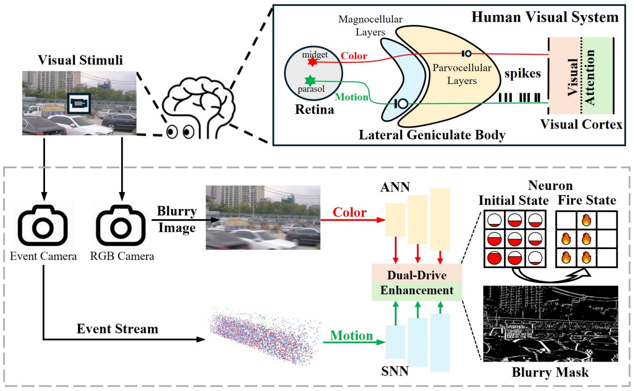


Figure 1. The working mechanism of human visual system after receiving visual stimuli and the proposed bio-inspired dual-drive hybrid network. The visual attention in visual cortex consists of the neuron-based (Pink) and synapse-based attention (Green) [16]. More details are provided in the supplementary material.

Drawing inspiration from biological systems, event-based cameras introduce an innovative paradigm for visual data acquisition [4, 14]. Event camera captures changes in brightness in high temporal resolution that naturally emphasize high-contrast edges. The explicit contrast edge information and the implicit temporal correlation among the event streams help to recover details lost in blurry images [38, 22]. Effective integration of event streams and frame-based images requires the precise extraction of motion features from event streams and background details from blurry images. However, mainstream ANN methods integrate event streams into frame-based or voxel-based representations and process in different channels, losing the temporal dependency [23].

Spiking Neural Networks (SNNs)[3] are inherently suitable to handle asynchronous event streams. Visual systems integrating SNNs with event cameras demonstrate promising performance in addressing complex visual tasks, attributed to SNNs' capability to effectively preserve the temporal dependencies inherent in event data [25, 5, 13, 22]. Nevertheless, image restoration is distinct from other high-level tasks that cannot be entirely based on SNNs for feature extraction. This is because SNNs encode information via binary spike sequences [30], which are insufficient for preserving essential high bit details such as color and structure information. This limitation affects the pixel-level precision required for image restoration. For this reason, hybrid networks with Artificial Neural Networks (ANNs) for color information processing and Spiking Neural Networks (SNNs) for event processing are adopted, aiming to combine the advantages of both ANN and SNN [13, 5, 23].

To effectively integrate data from both sensor types, previous studies [27, 6, 36] have concentrated on the development of cross-modal attention mechanisms, achieving substantial performance improvements. However, previous synaptic-level attention mechanisms that modify weight magnitudes overlook the inherent physical differences between sensor modalities, limiting the effectiveness. This limitation arises from two main factors: Firstly, the distribution of events is non-uniform. In the blurry regions, events generated by low-contrast scenes exhibit a sparse distribution. This sparsity is insufficient to trigger neuron responses in SNNs, resulting in the ineffectiveness of cross-modal attention mechanisms. Second, event features exhibit redundancy. Within the exposure duration, numerous events are triggered by repetitive movements associated with the same objects or the contours that are the details typically lost in blurry images. Existing approaches that utilize all event features for cross-modal attention fail to effectively discriminate the specific motion features responsible for the blurriness in these areas, leading to suboptimal performance in motion deblurring.

To better leverage the multi-modal information for Event-based Motion Deblurring, this paper introduces a Bioinspired Dual-Drive Hybrid Network (BDHNet). Specifically, we designed a Neuron Configurator Module (NCM) as visual enhancement from image data to event data, to improve the performance in blurry regions with sparse events. The NCM module utilizes image features for cross-modal initialization of the neurons' membrane potential and threshold in the SNN blocks, enabling a pixel-level dynamic adjustment. Such behavior is categorized as neuron-based attention, which is also present in the human visual system as the baseline increase in neural activity that elevates neuron activity across specific visual areas [16]. The configuration enables more neural responses in the blurry regions when the event stream is sparse, and significantly enhances the SNN's capacity to extract detailed motion features. For the precise motion clue extraction, we introduce a Region of Blurry Attention module (RBAM) to enhance the synapse-based attention as visual enhancement from event data to image data. The RBAM module integrates localized spike data with image features to generate a mask specifically targeting blurry regions. This mask is subsequently utilized to selectively recalibrate the cross-modal feature fusion, strategically concentrating the network's perceptual focus on blurry areas, thereby enhancing the deblurring performance. The main contributions of our work are summarized as follows:

- We propose a bioinspired dual-drive hybrid network with the neuron-based and enhanced synapse-based attention to mimic the visual attention capability of the human visual system.
- We introduce a Neuron Configurator Module for the dynamic configurations of the SNN neurons and a Region of Blurry Attention Module that creates a targeted blurry mask to facilitate cross-modal feature fusion.
- Subjective and objective evaluations demonstrate that our BDHNet has achieved SOTA in varying blurry conditions in GoPro, REBlur and MS-RBD datasets.

2. Background and Related Work

2.1. Event-based Motion Deblurring

Event cameras capture continuous motion data with low latency, providing vital cues for enhancing motion deblurring. Recent researches have achieved notable advancements and demonstrate the effectiveness of integrating event data. EDI [24] establishes a rigorous mathematical integration between blurry images, event data, and sharp reference frames. eSL-Net [32] applies sparse learning to simultaneously denoise and enhance the resolution of images shaped from event data, effectively restoring high-quality results. EVDI [38] presents a comprehensive framework for event-based motion deblurring and frame interpolation, utilizing the low latency of event cameras to mitigate motion blur and enhance frame prediction. DS-Deblur [35] implements a dual-stream architecture that combines adaptive feature fusion with recurrent spatio-temporal transformations, refining image clarity. GEM [39] introduces a scale-aware network that adjusts to varying spatial and temporal scales, employing a self-supervised learning strategy to adapt to diverse real-world scenarios. MTGNet [22] proposes multi-temporal granularity network that efficiently merges voxel-based and point cloud-based events to optimize the exploitation of the inherent high temporal resolution.

Recent advancements leverage cross-modal attention mechanisms for effective multi-modal integration, achieving notable enhancements. EFNet [27] incorporates a multi-head attention mechanism to integrate data across different modalities. EIFNet [34] improves motion deblurring by efficiently processing both unique and shared features through a dual cross-attention mechanism, enhancing feature integration and differentiation. MAENet [28] utilizes alignment and multi-head attention to coherently fuse features, reducing inter-modal inconsistencies. STC-Net [36] implements differential-modality calibration and co-attention to enhance spatial fusion and model cross-temporal dependencies between frames and events using motion information.

Despite significant progress in event-based image deblurring, current methodologies exhibit fundamental limitations. Firstly, attention mechanisms at the synaptic level that only adjust the weight magnitudes are insufficient limited to the non-uniform distribution and inherent redundancy of event data. Moreover, conventional CNN-based approaches fail to adequately preserve the intrinsic temporal dependencies of event data, reducing the overall effectiveness of the deblurring process.

2.2. Spiking Neural Networks

Spiking Neural Networks, as bioinspired computational frameworks, are inherently suited to handle the asynchronous and sparse characteristics of event data [10, 15]. The most common neuron model is the Leaky Integrate-and-Fire (LIF) model with iterative expression [33]. At each timestep t, the neurons in the l-th layer integrate the postsynaptic current $\boldsymbol{c}^l[t]$ with previous membrane potential $\boldsymbol{u}^l[t-1]$, the mathematic expression is illustrated in Equation (1):

$$\boldsymbol{u}^l[t] = (1 - \frac{1}{\tau})\boldsymbol{u}^l[t-1] + \boldsymbol{c}^l[t], \tag{1}$$

where τ is the membrane time constant. $\tau > 1$ as the discrete step size is 1. The postsynaptic current $\boldsymbol{c}^l[t] = \mathcal{W}^l * \boldsymbol{s}^{l-1}[t]$ is calculated as the product of weights $\mathcal{W}^l$ and spikes from the preceding layer $\boldsymbol{s}^{l-1}[t]$, simulating synaptic functionality, with $*$ indicating either a fully connected or convolutional synaptic operation.

Neurons produce spikes $\boldsymbol{s}^l[t]$ via the Heaviside function Θ when the membrane potential $\boldsymbol{u}^l[t]$ surpasses the threshold V_{th}, as depicted in Equation (2):

$$\boldsymbol{s}^l[t] = \Theta(\boldsymbol{u}^l[t] - V_{\mathrm{th}}) = \begin{cases} 1, & \text{if } \boldsymbol{u}^l[t] \geq V_{\mathrm{th}} \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

After the spike, the neuron updates the membrane potential $\boldsymbol{u}^l[t]$ according to the reset mechanism as shown in Equation (3):

$$\boldsymbol{u}^l[t] = \boldsymbol{u}^l[t] - V_{\mathrm{th}}\boldsymbol{s}^l[t], \tag{3}$$

where the $V_{\mathrm{th}} \in \mathbb{R}$ is generally a global scalar that controls the firing and reset process for the neurons in each layers.

2.2.1 SNN-based Image Restoration

Recent works leverage SNNs for effective multi-modal image restoration, achieving impressive results. EMFHNet [21] introduces an event-enhanced multi-modal fusion hybrid network, incorporating an SNN encoder to efficiently process and denoise event data. SC-Net [5] effectively combines SNNs and CNNs to exploit the sparse temporal and spatial characteristics of event stream, enhancing event-driven video restoration. ESDNet [26] designs spiking residual block and attention mechanisms to enhance image deraining, effectively addressing the challenges of binary activation and complex training dynamics. EDHNet [13] introduces a hybrid event-driven network with a bimodal fusion module to effectively identify and remove rain streaks, significantly improving video deraining. Motion-SNN [23] employs a spiking neural network and a hybrid feature extraction encoder to optimize event-based image deblurring, seamlessly merging high-temporal-resolution event data with the image data for enhanced clarity.

However, current spike-based image restoration methods, with the uniform neuron configurations in the SNN branch, lack adaptability to the non-uniform distribution of event data and fail to harness complementary multi-modal inputs, resulting in compromised performance.

3. Method

3.1. Problem Formulation

The event-based motion deblurring network is informed by the human visual system, which efficiently manage complex environmental conditions. This capability is reflected in the network's design, where visual stimuli are methodically broken down and processed in a hierarchical and parallel manner as color and motion, as shown in Figure 1.

In the motion deblurring task, the conventional RGB camera captures the color and texture details of the scenario, and event camera provides the motion information. The blur accumulation process can be modeled by the intensity of sharp images $\mathcal{I}(t)$ as:

$$\mathcal{B} = \frac{1}{T}\int_{f-T/2}^{f+T/2} \mathcal{I}(t)\,dt. \tag{4}$$

where $\mathcal{B}$ denotes the blurry image, f indicates the latent time stamp of the sharp image and T is the exposure period of the sensors.

For the bioinspired event camera, events are emitted asynchronously each time the log-scale brightness change exceeds the positive event threshold $c > 0$:

$$\log(\mathcal{I}(t,x)) - \log(\mathcal{I}(f,x)) = p \cdot c, \tag{5}$$

where $\log(\mathcal{I}(t,x))$ and $\log(\mathcal{I}(f,x))$ denote the log-scale intensity of pixel x at time t and f, and p is the polarity of event data.

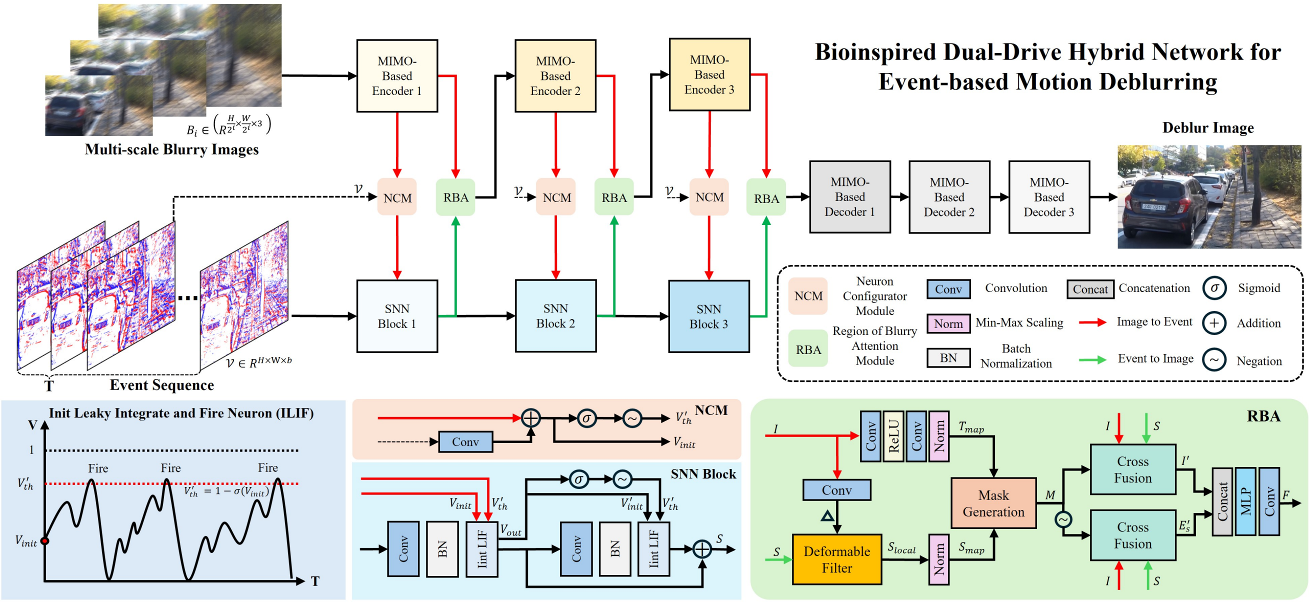


Figure 2. The overall framework of BDHNet. The event stream is shaped into the voxel-based representation $\mathcal{V}$. B_i are the multi-scale blurry images. I and S denote the image and spike features respectively. V_{init} and V'_{th} are the initialized membrane potential and threshold. T_{map} and S_{map} stand for the local spike map and threshold map for the blurry mask generation. M is the region of blurry mask. I' and E'_s are the image and event features after cross attention. The modules of the same name are shaded darker to indicate deeper network layers.

$\mathcal{I}(t)$ can be computed based on the events generated by the current pixel within the exposure period $\forall t \in T$ as:

$$\mathcal{I}(t) = \mathcal{I}(f) \cdot \exp\left(c \cdot \int_f^t p(s)\,ds\right), \tag{6}$$

where $\mathcal{I}(f)$ is the latent sharp image.

Substitute Equation (6) into Equation (4), we deduce the following formula:

$$\mathcal{I}(f) = \frac{\mathcal{B}}{\frac{1}{T}\int_{f-T/2}^{f+T/2} \exp\left(c\int_f^t p(s)\,ds\right)\,dt}, \tag{7}$$

Since the direct restoration of $\mathcal{I}(f)$ via Equation (7) often faces challenges due to the instability of event threshold c, learning-based methods are employed to more accurately model the statistical characteristics of events $\mathcal{E}$ as:

$$\mathcal{I}(f) = \mathrm{Deblur}(f; \mathcal{B}, \mathcal{E}), \quad \forall f \in T, \tag{8}$$

where $\mathrm{Deblur}(\cdot)$ indicates a motion deblurring network.

3.2. Network Architecture

The overall framework of our proposed Bioinspired Dual-Drive Hybrid Network is shown in Figure 2. We adopt a classical encoder-decoder architecture for our approach. Initially, the multi-scale blurry images $\mathcal{B} \in (\mathbb{R}^{H\times W\times 3}, \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 3}, \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 3})$ are fed into the ANN-based image branch, where we use the MIMO-Based Encoder [9] as our fundamental block, to extract relevant features. Simultaneously, the corresponding event stream is shaped into the voxel-based representation $\mathcal{E} \in \mathbb{R}^{H\times W\times b}$ and fed into the SNN-based event branch, which facilitates the extraction of motion features. Following each layer of the two branches, the Dual-Drive Enhancement is employed to mimic the visual attention in the human visual system. It consists of a Neuron Configurator Module (NCM) for dynamic setting the neuron configurations and a Region of Blurry Attention Module (RBAM) that strategically focuses on motion features causing blurry effects to enhance cross-modal feature fusion. In the decoder, the MIMO-Based Decoder is applied for the image reconstruction and the PSNR Loss function [8] is applied to precisely optimize the network's parameters for optimal performance.

3.2.1 Neuron Configurator Module

The Neuron Configurator Module (NCM) is designed for visual enhancement from image data to event data. It strategically modulates neuronal responses to concentrate on critical regions as the neuron-based attention. Unlike previous approach [1] that initialize the neuron's membrane potential at the first timestep, NCM employs image features to set both the initial timestep membrane potentials and the neuron thresholds across all timesteps based on input stimuli characteristics, enabling a pixel-level dynamic adjustment. Specifically, the structural and chromatic features, initially extracted from the blurry images $\mathcal{B}$, are utilized to identify the blurry regions. These features are then integrated with the attributes extracted from event data $\mathcal{E}$ through a shallow convolution layer, designed to mitigate domain discrepancies between the two modalities.

The initial membrane potential V_{init} at $t = 0$ is set based on the integrated features from blurry image $\mathcal{B}$ and event data $\mathcal{E}$ as:

$$V_{\text{init}} = \phi_{\text{init}}(\mathcal{B}) + \psi_{\text{init}}(\mathcal{E}), \tag{9}$$

where ϕ_{init} is the feature encoder of the blurry image and ψ_{init} is the shallow convolution layer of the event data.

To fully elevates neuron activity across blurry areas, the threshold V'_{th} is redesigned according to the initial membrane potential as:

$$V'_{\text{th}} = 1 - \sigma(V_{\text{init}}), \tag{10}$$

where σ is the Sigmoid function, which rescales the initial feature to the range of 0 to 1. Unlike the global scalar threshold $V_{\text{th}} \in \mathbb{R}$ in vanilla LIF in Equation (2) and (3), the threshold $V'_{\text{th}} \in \mathbb{R}^{H\times W\times C}$ has the same dimension with the feature map, providing more fine-grained control over the reset and firing processes of the neurons in the same layer.

After the neuron configuration, the membrane potential update formula of each neuron is simplified to utilize V_{init} to directly set the membrane potential at the initial timestep, with subsequent timesteps updated as follows:

$$\boldsymbol{u}[t] = \begin{cases} V_{\text{init}} & \text{if } t = 0 \\ (1 - \frac{1}{\tau})\boldsymbol{u}[t-1] + \boldsymbol{c}[t] & \text{otherwise} \end{cases}. \tag{11}$$

The condition for spike emission and the membrane potential reset strategy are as follows:

$$\boldsymbol{s}[t] = \begin{cases} 1, & \text{if } \boldsymbol{u}[t] \geq V'_{\text{th}} \\ 0, & \text{otherwise} \end{cases}. \tag{12}$$

After the spike, the soft reset process is updated as follows:

$$\boldsymbol{u}^l[t] = \boldsymbol{u}^l[t] - V'_{\text{th}} \odot \boldsymbol{s}^l[t]. \tag{13}$$

In the event data feature extraction branch, we have implemented an SNN Block with residual connections as shown in Figure 2. This block comprises two layers: the first layer's neurons are dynamically configured at the pixel level with the cross-modal initialization. The membrane potentials from the first layer subsequently inform the initialization and configuration of the second layer. Outputs from both layers are fused through a residual connection, culminating in the final output spikes S. This configuration, facilitated by the Neuron Configurator Module, enables precise pixel-level adjustments of neuronal activity, enhancing the neurons' responsiveness in blurry regions and thus improving the extraction of motion features from event data.

3.2.2 Region of Blurry Attention Module

The Region of Blurry Attention Module (RBAM) is proposed for visual enhancement from event data to image data as shown in Figure 2. The RBAM module capitalizes on image features I from the ANN branch and spike features S from the SNN branch to generate a mask M delineating blurry regions. This mask serves to capture accurate motion clues from event features, thereby refining the cross-modal feature fusion.

Specifically, spikes activated by multiple pixels within the SNN branch originate from the same moving object in the scenario, with pertinent information about the motion contours embedded in the blurry image. Accordingly, the RBAM employs a deformable filter for localized aggregation of spike features.

Within this filter, the spike features are firstly integrated in the temporal dimension as,

$$S_{sum} = \sum_t S(x, y, t). \tag{14}$$

The bias for each deformable convolution kernel associated with individual pixels is determined by image features processed through a shallow convolution layer and the deformable filter is formulated as,

$$S_{local} = \text{DeformSum}(S_{sum}, \text{Conv}(I)), \tag{15}$$

Subsequent to the aggregation by the deformable filter, the spike map S_{map} undergoes min-max normalization to facilitate uniformity as follows:

$$S_{map} = \text{Norm}(S_{local}), \tag{16}$$

where Norm is the Min-Max scaling operation as $\text{Norm}(X) = \frac{X - \min(X)}{\max(X) - \min(X)}$.

Further, a pixel-level threshold map T_{map} is generated through the image features as,

$$T_{map} = \text{Norm}(\text{Conv}(\text{ReLU}(\text{Conv}(I)))), \tag{17}$$

this map is then subjected to pixel-level binarization against the spike map, producing a mask that identifies the blurry regions as follows,

$$M = \begin{cases} 1 & \text{if } S_{map} \geq T_{map} \\ 0 & \text{otherwise} \end{cases}. \tag{18}$$

This unsupervised identification process for blurry areas significantly enhances cross-modal feature fusion by utilizing this dynamically generated mask.

For the mask-guided fusion, the spike features undergo a temporal convolution to adaptively integrate the temporal information and obtain the event feature E_s. The image

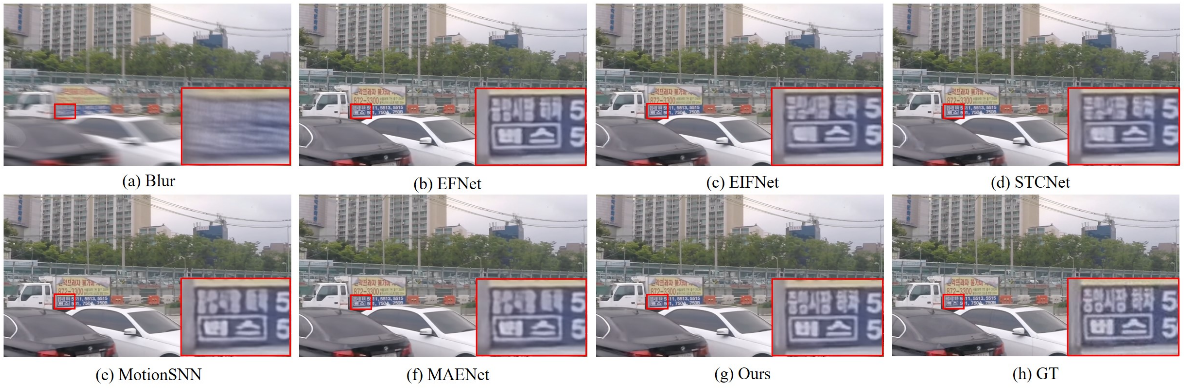


Figure 3. Qualitative comparisons under GoPro dataset. Best viewed on a screen and zoomed in.

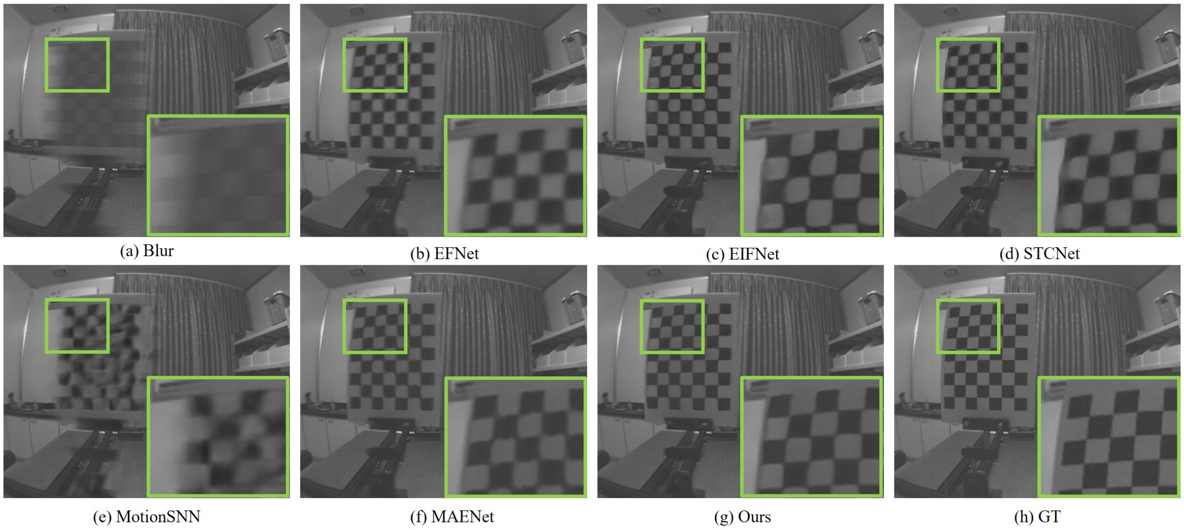


Figure 4. Qualitative comparisons under REBlur dataset. Best viewed on a screen and zoomed in.

feature I and the event feature E_s are applied with the cross attention operation as,

$$\begin{cases} I' = I + \text{mask} \cdot \text{Attention}(Q_I, K_{E_s}, V_{E_s}), \\ {E_s}' = E_s + (1 - \text{mask}) \cdot \text{Attention}(Q_{E_s}, K_I, V_I), \end{cases} \tag{19}$$

where Attention denotes multi-head attention operation.

The fusion features are generated through the channel-wise concatenation as,

$$F = \text{Conv}(\text{MLP}(\text{Concat}(I', {E_s}'))). \tag{20}$$

The final deblur images are reconstructed through the MIMO-based Decoder [9] and the entire training process is conducted in an end-to-end fashion.

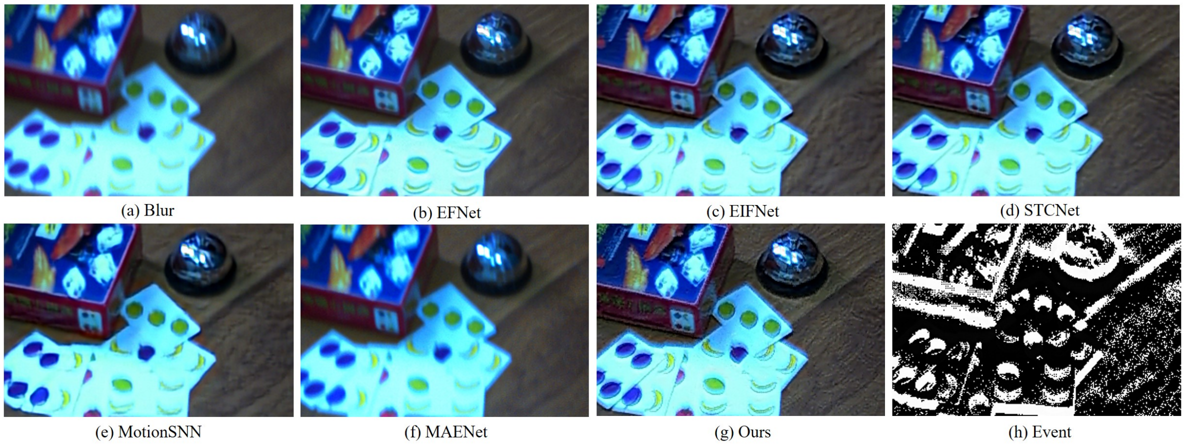
4. Experiment

4.1. Datasets

We evaluate the proposed method with GoPro, REBlur and MS-RBD datasets containing both synthetic and real-world scenarios.

GoPro: We evaluate the deblurring performance on GoPro dataset [27], which is the benchmark dataset for the image motion deblurring. It consists of 3214 pairs of blurry and sharp images, with 2103 pairs for training and 1111 pairs for testing. The resolution of all images is 1280×720 and the blurry images are produced by averaging several adjacent high-speed sharp images. The event data is generated through the ESIM simulator. In this work, the raw event data is shaped into voxel-based representation for each image following EIFNet and the timestep in $\mathcal{V}$ is set to $b = 12$.

REBlur: REBlur dataset [27], captured by DAVIS for



(a) Blur (b) EFNet (c) EIFNet (d) STCNet

(e) MotionSNN (f) MAENet (g) Ours (h) Event

Figure 5. Qualitative comparisons under MS-RBD dataset. Best viewed on a screen and zoomed in.

Table 1. Performance comparison on GoPro and REBlur datasets with and without fine-tune. The best results are in bold.

Method	Input		GoPro		REBlur		REBlur w/o Fine-tune	
	Image	Events	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
SRN [29]	✓	×	30.26	0.934	35.10	0.961	\	\
HINet [8]	✓	×	32.71	0.959	35.58	0.965	\	\
NAFNet [7]	✓	×	33.69	0.967	35.48	0.962	\	\
Restormer [37]	✓	×	32.92	0.961	35.50	0.959	\	\
MSDI-Net [18]	✓	×	33.28	0.964	36.14	0.968	\	\
UFPNet [11]	✓	×	34.06	0.968	36.11	0.968	\	\
EFNet [27]	✓	✓	35.46	0.972	38.02	0.975	27.43	0.898
MotionSNN [23]	✓	✓	35.18	0.971	36.32	0.968	34.63	0.957
EIFNet [34]	✓	✓	35.99	0.973	37.81	0.976	35.75	0.965
STCNet [36]	✓	✓	36.45	0.975	37.78	0.976	35.64	0.966
MAENet [28]	✓	✓	36.07	0.976	38.46	0.978	32.86	0.950
Ours	✓	✓	**37.04**	**0.977**	**38.50**	**0.978**	**36.01**	**0.967**

real event-based motion deblurring, comprises 1,389 image pairs with 486 designated for training and 903 for testing. It contains diverse linear and nonlinear indoor motions. Each image has a resolution of 260×360, consisting of real-world event data along with the corresponding blurry and sharp images.

MS-RBD: MS-RBD dataset [39] is the multi-scale blurry dataset captured in the real-world scenario. The dataset contains 32 sequences of data with 22 indoor and 10 outdoor scenes. The resolution of all images is 288×192 with the corresponding events. We evaluate the deblurring performance on MS-RBD with a focus on the generalization ability in the real-world scenes, where the blur caused by camera ego-motion and dynamic scenes.

4.2. Implementation Details

During the training process, we deploy our proposed network in the PyTorch framework on a single NVIDIA RTX 4090 GPU. The ADAM optimizer [17] is utilized with an initial learning rate of 1×10^{-4}, which is scheduled to decrease at the 60th and 80th epochs, over a total of 120 epochs. For data augmentation, horizontal and vertical flipping, rotation, and random crop are applied. The crop size is set to 512 for the GoPro dataset. Fine-tuning on the REBlur dataset is conducted over 30 epochs with an initial learning rate of 1×10^{-5}. The crop size for REBlur is set to 256 and other configurations are kept the same as for GoPro. Our evaluation metrics include PSNR and SSIM.

4.3. Comparison Experiments

We compare our proposed BDHNet to SOTA image-only and event-based deblurring methods on GoPro, REBlur and MS-RBD datasets for a comprehensive evaluation. The comparison methods include image-only method: SRN [29], HINet [8], NAFNet [7], Restormer [37], MSDI-Net [18], and UFPNet [11]. The event-based methods consists of EFNet [27], MotionSNN [23], EIFNet [34], STCNet [36], and MAENet [28]. The event-based methods are

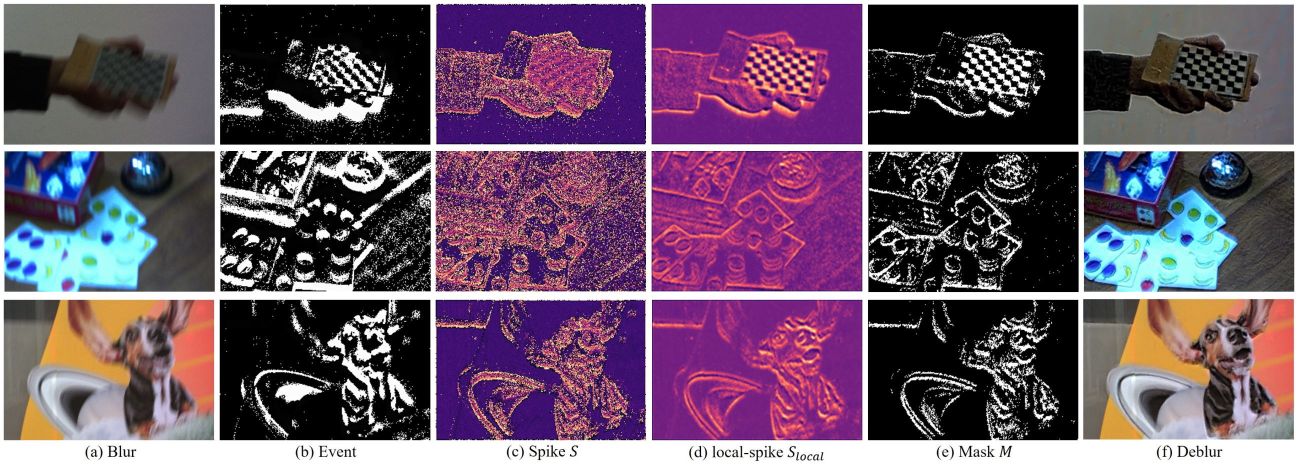

(a) Blur
(b) Event
(c) Spike S
(d) local-spike S_{local}
(e) Mask M
(f) Deblur

Figure 6. Visualization of the unsupervised blurry mask generation process under MS-RBD dataset.

Table 2. Ablation study of the proposed method on GoPro dataset. Image means input blurry image, Event stands for the corresponding event data, NCM is the Neuron Configurator Module, Mask is the region of blurry mask in RBAM, CA means cross attention, add is the addition operation. The best results are in bold.

Image	Event	NCM	Mask	Fusion Module	PSNR / SSIM
✓	×	×	×	×	31.64 / 0.949
✓	✓	×	×	Add	36.29 / 0.972
✓	✓	✓	×	Add	36.51 / 0.973
✓	✓	✓	×	CA	36.60 / 0.974
✓	✓	✓	✓	Add	36.84 / 0.975
✓	✓	✓	✓	CA	**37.04 / 0.977**

all based on the raw event data produced by EFNet and we utilize the open-source checkpoint to evaluate the performance on GoPro dataset. For a fair comparison, all methods are trained under the optimal parameter settings as specified in the respective papers if there are no open-source checkpoint in REBlur dataset. Our comparison metrics follow the benchmark established by EFNet and MAENet, maintaining consistency in metric calculations libraries.

Table 1 provides a detailed comparative analysis of the comparison deblurring methods evaluated on the GoPro and REBlur datasets. Notably, our method significantly outperforms others in both datasets, achieving the highest performance metrics with a PSNR of 37.04 and an SSIM of 0.977 on the GoPro dataset, and a PSNR of 38.50 and an SSIM of 0.978 on the REBlur dataset. These results underscore our BDHNet's superior ability to mitigate blur effects under varied conditions.

Specifically, the performance on the REBlur dataset without fine-tuning is particularly noteworthy. Our method, when applied directly without specific adaptation to the REBlur dataset (trained solely on the GoPro data), achieves the best performance with a PSNR of 36.01 and an SSIM of 0.967. The robust generalization ability of our model derives from its bio-inspired architecture, which emulates the visual attention mechanism intrinsic to the human visual system. This design enables adaptive modulation of neuron responses, enhancing the model's focus on blurry regions. The neuron-based attention allows for effective adjustment to various blur intensities encountered across different datasets, obviating the need for dataset-specific tuning. This capability not only ensures consistent performance under diverse imaging conditions but also underscores the model's superiority for practical applications in real-world scenarios.

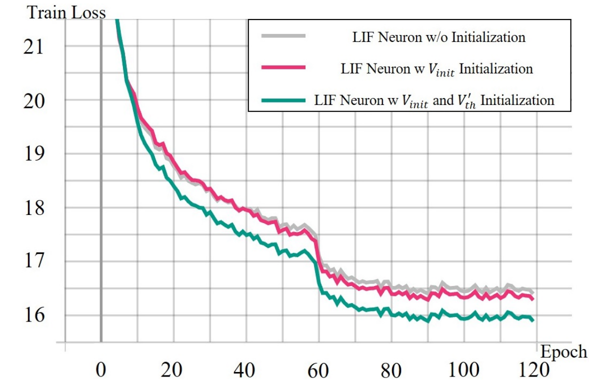
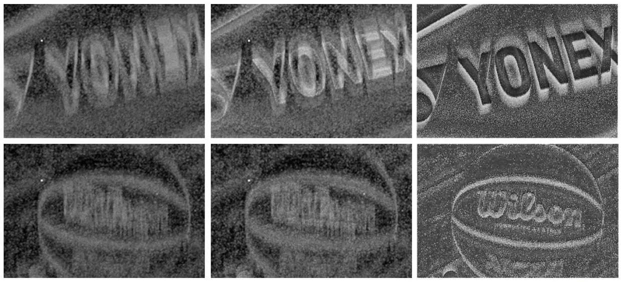


Figure 7. Training loss under different neuron configurations.

The visual comparisons presented in Figure 3, Figure 4, and Figure 5 effectively demonstrate the superior performance of our method in deblurring tasks, evidencing enhanced detail recovery and reduced spatial distortions across various scenarios. Our model consistently outperforms competing methods, achieving clearer and more precise reconstructions. Specifically, it excels in restoring sharper text in Figure 3 and finer structural details in Fig-



(a) LIF w/o Initialization (b) LIF w V_{init} (c) LIF w V_{init} and V'_{th}

Figure 8. Visualization of the neuron responses of different configurations under MS-RBD dataset.

ure 4 and Figure 5, significantly improving legibility and image quality. This showcases the effectiveness of our BDHNet in accurately perceiving and processing motion information, which is crucial for high-quality motion deblurring in practical applications. More visual comparison results are provided in the supplementary material.

The objective metrics and subjective evaluations in our study highlight our method's superior performance and exceptional generalization ability across diverse datasets.

4.4. Ablation Studies

To evaluate the effectiveness of the key components in the proposed BDHNet, we conduct comprehensive ablation studies on the GoPro datasets, as shown in Table 2. We also visualize the generation process of the blurry region mask step by step in Figure 6.

Effectiveness of the Neuron Configurator Module. The comparative analysis between the second and third rows of Table 2 substantiates the efficacy of the Neuron Configurator Module as the neuron-based attention. Compared to the deblurring performance under conditions without initial neuron configuration, the PSNR increased by 0.22 dB. Our evaluation includes a focused comparison between methods that initialize only the membrane potential at the first timestep and our proposed approach, which encompasses the initialization of membrane potential and the dynamic configuration of neuron thresholds across all timesteps. As evidenced by the visual comparisons in Figure 7, our method demonstrates superior training convergence. In contrast, initializing only the first timestep membrane potential [1] offers marginal improvements under conditions without configuration.

Figure 8 further validates the effectiveness of our configuration method by visualizing the spike outputs produced by neurons under three different settings for the same scene. It illustrates that our method, which leverages bio-inspired visual attention mechanisms from the human visual system, effectively modulates spikes triggered by varying visual stimuli through dynamic neuron configuration. This method proficiently concentrates spikes on blurry regions or motion-inducing edges. In contrast, configurations that either only initialize the initial membrane potential or without initialization struggle to capture motion characteristics effectively, thus yielding suboptimal deblurring results.

Effectiveness of the Region of Blurry Attention Module. In the RBAM module, our architecture is divided into two principal components. The first component focuses on generating a mask for blurry regions in an unsupervised manner that utilizes both spike and image features. The second component applies the mask to guide the cross-modal feature fusion. The efficacy of this masking process is substantiated by incremental improvements in PSNR of 0.33 dB and 0.44 dB, as shown in the comparative analysis between rows 3 and 5, and rows 4 and 6 in Table 2.

The mechanics of this unsupervised mask generation are detailed through the visualizations in Figure 6, based entirely on the real-world MS-RBD dataset, thus demonstrating the robustness and the generalization ability of our approach. Column c and d of Figure 6 utilize heatmaps to visually demonstrate that spike features, post-processing with a deformable filter, are more accurately focused on the blurry areas or edges inducing blur, as illustrated in column d. This enhanced focus is facilitated by the deformable bias introduced by image features, which implicit captures motion information. The resulting masks, generated from the amalgamation of local aggregation spike maps and threshold maps derived from image features as depicted in column e, effectively delineate the regions of blurriness. The region of blurry mask accurately captures the motion clues that cause the blurry effects from the event features, thereby guiding the effective cross-modal feature fusion.

We further evaluate the effectiveness of the generated masks with various feature fusion methods: pixel-level addition and cross-modal attention. According to the data presented in row 5 of Table 2, our approach utilize addition for fusion exhibited significant deblurring capabilities. Subsequently, integrating a cross attention fusion strategy further augmented our model's performance, enabling it to reach state-of-the-art level in motion deblurring.

5. Conclusion

In this paper, we propose the Bio-inspired Dual-Drive Hybrid Network (BDHNet) for event-based motion deblurring. Drawing inspiration from the human visual system, the dual-drive enhancement strategy effectively mitigates the impact of blur resulting from camera or scene motion. The integration of the Neuron Configurator Module (NCM) and the Region of Blurry Attention Module (RBAM) enables dynamic and precise adaptation to blurry areas. Comprehensive evaluations demonstrate that BDHNet sets a new standard in the field, surpassing existing technologies in both synthetic and real-world scenarios.

References

[1] Asude Aydin, Mathias Gehrig, Daniel Gehrig, and Davide Scaramuzza. A hybrid ann-snn architecture for low-power and low-latency visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5701–5711, 2024.

[2] Yuval Bahat, Netalee Efrat, and Michal Irani. Non-uniform blind deblurring by reblurring. In *Proceedings of the IEEE international conference on computer vision*, pages 3286–3294, 2017.

[3] Maxence Bouvier, Alexandre Valentian, Thomas Mesquida, Francois Rummens, Marina Reyboz, Elisa Vianello, and Edith Beigne. Spiking neural networks hardware implementations and challenges: A survey. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 15(2):1–35, 2019.

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

[5] Chengzhi Cao, Xueyang Fu, Yurui Zhu, Zhijing Sun, and Zheng-Jun Zha. Event-driven video restoration with spiking-convolutional architecture. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[6] Kang Chen and Lei Yu. Motion deblur by learning residual from events. *IEEE Transactions on Multimedia*, 2024.

[7] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022.

[8] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 182–192, 2021.

[9] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4641–4650, 2021.

[10] Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks. *Advances in Neural Information Processing Systems*, 34:21056–21069, 2021.

[11] Zhenxuan Fang, Fangfang Wu, Weisheng Dong, Xin Li, Jinjian Wu, and Guangming Shi. Self-supervised non-uniform kernel estimation with flow-based motion prior for blind image deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18105–18114, 2023.

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

[13] Xueyang Fu, Chengzhi Cao, Senyan Xu, Fanrui Zhang, Kunyu Wang, and Zheng-Jun Zha. Event-driven heterogeneous network for video deraining. *International Journal of Computer Vision*, pages 1–21, 2024.

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

[15] Yulong Huang, LIN Xiaopeng, Hongwei Ren, FU Haotian, Yue Zhou, LIU Zunchang, Bojun Cheng, et al. Clif: Complementary leaky integrate-and-fire neuron for spiking neural networks. In *Forty-first International Conference on Machine Learning*.

[16] Nancy Kanwisher and Ewa Wojciulik. Visual attention: insights from brain imaging. *Nature reviews neuroscience*, 1(2):91–100, 2000.

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

[18] Dasong Li, Yi Zhang, Ka Chun Cheung, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Learning degradation representations for image deblurring. In *European conference on computer vision*, pages 736–753. Springer, 2022.

[19] Haoying Li, Ziran Zhang, Tingting Jiang, Peng Luo, Huajun Feng, and Zhihai Xu. Real-world deep local motion deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1314–1322, 2023.

[20] Haoying Li, Jixin Zhao, Shangchen Zhou, Huajun Feng, Chongyi Li, and Chen Change Loy. Adaptive window pruning for efficient local motion deblurring. In *The Twelfth International Conference on Learning Representations*, 2024.

[21] Si-Qi Li, Yue Gao, and Qiong-Hai Dai. Image de-occlusion via event-enhanced multi-modal fusion hybrid network. *Machine Intelligence Research*, 19(4):307–318, 2022.

[22] Xiaopeng Lin, Hongwei Ren, Yulong Huang, Zunchang Liu, Yue Zhou, Haotian Fu, Biao Pan, and Bojun Cheng. Event-based motion deblurring via multi-temporal granularity fusion. *arXiv preprint arXiv:2412.11866*, 2024.

[23] Zhaoxin Liu, Jinjian Wu, Guangming Shi, Wen Yang, Weisheng Dong, and Qinghang Zhao. Motion-oriented hybrid spiking neural networks for event-based motion deblurring. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829, 2019.

[25] Hongwei Ren, Yue Zhou, Yulong Huang, Haotian Fu, Xiaopeng Lin, Jie Song, and Bojun Cheng. Spikepoint: An efficient point-based spiking neural network for event cameras action recognition. *arXiv preprint arXiv:2310.07189*, 2023.

[26] Tianyu Song, Guiyue Jin, Pengpeng Li, Kui Jiang, Xiang Chen, and Jiyu Jin. Learning a spiking neural network for efficient image deraining. *arXiv preprint arXiv:2405.06277*, 2024.

[27] 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 *European conference on computer vision*, pages 412–428. Springer, 2022.

[28] Zhijing Sun, Xueyang Fu, Longzhuo Huang, Aiping Liu, and Zheng-Jun Zha. Motion aware event representation-driven image deblurring. In *European Conference on Computer Vision*, pages 418–435. Springer, 2025.

[29] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8174–8182, 2018.

[30] Corinne Teeter, Ramakrishnan Iyer, Vilas Menon, Nathan Gouwens, David Feng, Jim Berg, Aaron Szafer, Nicholas Cain, Hongkui Zeng, Michael Hawrylycz, et al. Generalized leaky integrate-and-fire models classify multiple neuron types. *Nature communications*, 9(1):709, 2018.

[31] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *European conference on computer vision*, pages 146–162. Springer, 2022.

[32] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 155–171. Springer, 2020.

[33] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in neuroscience*, 12:331, 2018.

[34] Wen Yang, Jinjian Wu, Leida Li, Weisheng Dong, and Guangming Shi. Event-based motion deblurring with modality-aware decomposition and recomposition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8327–8335, 2023.

[35] Wen Yang, Jinjian Wu, Jupo Ma, Leida Li, Weisheng Dong, and Guangming Shi. Learning for motion deblurring with hybrid frames and events. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1396–1404, 2022.

[36] Wen Yang, Jinjian Wu, Jupo Ma, Leida Li, and Guangming Shi. Motion deblurring via spatial-temporal collaboration of frames and events. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6531–6539, 2024.

[37] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[38] Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17774, 2022.

[39] Xiang Zhang, Lei Yu, Wen Yang, Jianzhuang Liu, and Gui-Song Xia. Generalizing event-based motion deblurring in real-world scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10734–10744, 2023.