Spike-guided Motion Deblurring with Unknown Modal Spatiotemporal Alignment

Jiyuan Zhang[1,2] Shiyan Chen[1,2] Yajing Zheng[1,2*] Zhaofei Yu[1,2,3*] Tiejun Huang[1,2,3]
[1]School of Computer Science, Peking University
[2]National Key Laboratory for Multimedia Information Processing, Peking University
[3]Institute for Artificial Intelligence, Peking University
{jyzhang,2301112005}@stu.pku.edu.cn, {yj.zheng,yuzf12,tjhuang}@pku.edu.cn

Abstract

The traditional frame-based cameras that rely on exposure windows for imaging experience motion blur in high-speed scenarios. Frame-based deblurring methods lack reliable motion cues to restore sharp images under extreme blur conditions. The spike camera is a novel neuromorphic visual sensor that outputs spike streams with ultra-high temporal resolution. It can supplement the temporal information lost in traditional cameras and guide motion deblurring. However, in real-world scenarios, aligning discrete RGB images and continuous spike streams along both temporal and spatial axes is challenging due to the complexity of calibrating their coordinates, device displacements in vibrations, and time deviations. Misalignment of pixels leads to severe degradation of deblurring. We introduce the first framework for spike-guided motion deblurring without knowing the spatiotemporal alignment between spikes and images. To address the problem, we first propose a novel three-stage network containing a basic deblurring net, a carefully designed bi-directional deformable aligning module, and a flow-based multi-scale fusion net. Experimental results demonstrate that our approach can effectively guide the image deblurring with unknown alignment, surpassing the performance of other methods. Public project page: https://github.com/Leozhangjiyuan/UaSDN.

1. Introduction

The imaging principle of conventional frame-based cameras is based on the concept of exposure time windows. During the continuous imaging process, the time between the exposure windows of adjacent frames is unused. Due to the presence of exposure windows, when the camera is in high-speed motion or there are fast-moving objects in the scene, images may exhibit a blur effect. In recent years, many studies have focused on addressing the challenging task of recovering clear images from blurry ones utilizing deep-learning routes [4, 5, 16]. However, the texture or structural information in blurry images is partially lost during the exposure, making it difficult to infer clear textures directly from blurry images, as in Fig. 2 (Orange Box).

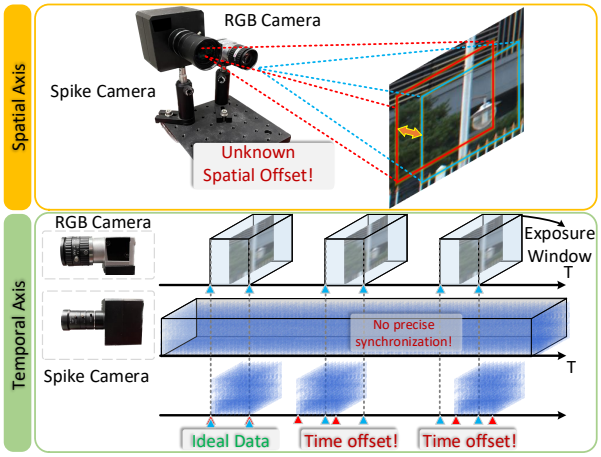

Figure 1. Illustration of unknown spatiotemporal alignment between a spike camera and an RGB camera.

Introducing spike cameras as clues of deblurring. Bio-inspired neuromorphic vision sensors have attracted increasing attention and achieved significant development. They possess advantages such as ultra-high temporal resolution, low latency, and high dynamic range. Due to the continuous recording of light intensity, neuromorphic cameras can be used to compensate for the lack of temporal information in image deblurring. Based on different sampling principles, they can be classified into event cameras [1, 21, 23] and spike cameras [12]. Event cameras asynchronously detect changes in light intensity and are highly sensitive to moving edges. Studies utilize events for image deblurring [28, 32]. **Spike cameras** accumulates

*Corresponding author.

photons and emits 20,000Hz spike streams. Due to the integration sampling principle, it can thoroughly record spatiotemporal information, including the texture and motion information of the scene. The precise texture information recorded in spikes can effectively guide the deblurring process. Our paper explores the effectiveness of using spike cameras to guide motion deblurring, uncovering the potential advantages of spike cameras in this novel task.

Take misalignment between cameras in real application into consideration. At present, most deblurring works based on event cameras assume that the imaging planes of the two cameras are absolutely coaxial [32, 33] and that the event stream and video are fully synchronized on the time axis. Such assumptions facilitate model design but are overly idealistic. In the real world, achieving strict alignment of a spike camera and an RGB camera in both spatial and temporal dimensions is quite challenging. As shown in Fig. 1, in this work, we consider the unknown alignment factors in the real world:

(A) *Spatial alignment of the two cameras is challenging.* Alignment of the coordinate systems of the two cameras can be achieved using methods such as beam splitters or manual calibration (e.g., chessboard calibration). However, beam splitters can result in significant loss of light, and differences in focal length, aperture, and relative positions between cameras, which can vary with the scene, make alignment operations complex and difficult.

(B) *Position displacements.* During motion, minor displacements in the relative positions of the two cameras occur due to vibrations. For current algorithms, pixel misalignment will lead to significant performance losses.

(C) *Synchronizing discrete image sequences with continuous spike streams on the time axis is challenging*. Technically, obtaining spikes strictly synchronized with blurry images requires acquiring the accurate exposure start and end times of the blurry images, which is complex. Moreover, in high-speed motion scenes, the two data streams are prone to small time deviations, leading to misalignment.

What we focus on and How we deal with the real-world misalignment? In response to the real-world misalignments mentioned above, we focus on investigating spike-guided motion deblurring with unknown spatiotemporal alignment between spikes and RGB images, as shown in Figs. 1 and 2. The proposed method allows for the use of two cameras simultaneously without the need for complex and rigorous spatial calibration and time alignment design. To address the problem, we propose an effective three-stage model **U**n**a**ligned **S**pike-guided **D**eblur **N**et (**UaSDN**). (1) In the first stage, we employ a simplified off-the-shelf network to perform basic deblurring. In this stage, we aim to leverage existing technology to alleviate the blurring and enhance textures. (2) In the second stage, we first use a shallow net to estimate coarse light intensity from the spike stream. Besides, we design a bidirectional deformable modal alignment module. The module achieves better alignment by learning bidirectional deformable convolutions between spike features and image features. (3) In the third stage, we design a multi-scale dense fusion network combining the optical flow, which performs feature fusion by aligning pixel motion offsets between modalities and fuse features in a multi-scale network. Through comparisons with various networks, experimental results demonstrate the effectiveness of our method in motion deblurring with unknown spatiotemporal alignment between modalities. To comprehensively validate the model's performance, we construct spike datasets for training and validation. Our contributions can be summarized as follows:

- We first explore the motion deblurring tasks guided by the spike camera and deal with the unknown spatiotemporal alignment between two modalities in the real world.
- We propose the three-stage UaSDN model which contains a bi-directional deformable modal alignment module and a flow-based fusion module.
- Experimental results demonstrate that our model significantly outperforms other methods on motion deblurring with unknown spatiotemporal alignment.

2. Related Work

2.1. Advances on Spike Cameras

The spike camera outputs spikes with an ultra-high temporal resolution of 20,000Hz [12]. Due to the sampling principle, it effectively captures texture in both moving and stationary areas, demonstrating potential in high-speed scenarios. Research focuses on utilizing spike cameras for image recovery or high-level visual tasks. Early studies involved direct statistics of spike counts or intervals for a rough estimation of light intensity [57]. Zhao *et al*. [52] introduces the first learning-based method. Zhang *et al*. [47] leverages wavelet transforms on spikes, and enhances the representational capability of spikes. Some bio-inspired works employ spiking neural networks (SNN) [53, 55] to process spikes. The introduction of self-supervised schemes enhanced the generalization capability [6]. Many applications have emerged, such as super-resolution reconstruction [51] and high-dynamic-range restoration [3], spike-guided video interpolation [41], denoising [11] and tracking [54].

2.2. Frame-based Motion Deblurring

In recent years, with the rise of deep learning, CNN-based approaches have shown remarkable performance in frame-based deblurring tasks. Popular technical routes is employing end-to-end training [4, 9, 16, 24, 34, 44]. Some works [17, 18] employ generative networks. Rather than mapping a blurry image one-to-one to a single sharp image, some work [2, 13, 26, 27, 31, 38, 56, 58] attempts to mine

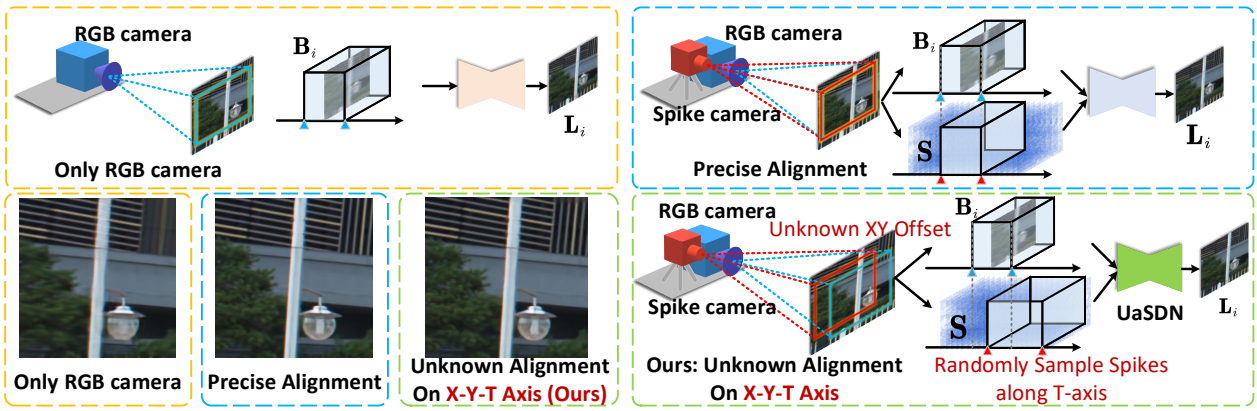

Figure 2. Illustration of the process of deblurring images in different scenarios. Orange box: deblurring directly from the RGB camera; blue box: deblurring with precise alignment of spike-RGB data; green box: unknown alignment in X-Y-T for spike-RGB data.

clues from video sequence input to assist in deblurring. However, video-based methods always use future frames, which is not applicable in the real world. Besides, many image-based image restoration models are able to deal with deblurring [4, 5, 19, 20, 39, 43–45].

2.3. Event-guided Motion Deblurring

As neuromorphic vision sensors, event cameras have been used in deblurring. Pan *et al*. [28] builds a computational model between blurred images and events. Several works [14, 36] utilize motion information and recurrent networks for aligning. Shang *et al*. [29] enhances the deblurring by finding adjacent sharp frames. Kim *et al*. [15] and Weng *et al*. [40] deal with images with unknown exposure time. EFNet [32] introduces an attention-based module to better fuse events. Cho *et al*. [8] proposed a solution for cameras with spatial offsets. Zhang *et al*. [49] proposes to generalize the deblurring ability to various spatial resolutions and blur levels. Some works [22, 28, 33, 37, 42, 48] combine the motion deblurring and frame interpolation guided by events. However, most studies assume that scenes captured by the event camera and the traditional camera are entirely identical, i.e. two camera coordinates coincide perfectly along the spatial axis and events have the same center and duration as the exposure time of blurred images on the temporal axis. Our study aims to address motion deblurring in real scenes unknown of the alignment between cameras in both time and space.

3. Method

3.1. Problem Analysis

We mainly focus on dealing with the motion deblurring in high-speed scenarios guided by a spike camera while *not knowing the alignment between the RGB camera and spike camera on both the spatial axis and temporal axis*. In this section, we give a specific analysis of how to utilize spike data for deburring spatiotemporal unaligned RGB frames.

An RGB camera outputs discrete RGB blurry frames $\mathbb{B} = \{\mathbf{B}_i | i \in \mathbf{N}\}$ with the exposure time of T_B, which is a shorter duration than the time duration between successive frames T_F (the frame rate $f = \frac{1}{T_F}$). A spike camera outputs the spike stream $\mathbf{S}$ with the rate of 20000Hz. Each pixel (x, y) of a spike camera on the $H_S \times W_S$ plane independently receives coming photons continuously at any timestamps t, converts the light signals into electrical current $I_{x,y}(t)$, and accumulates the voltage $V_{x,y}$. Whenever the voltage reaches the preset threshold Θ, the pixel fires a spike and resets the voltage to 0. The spike signals are read out with a very short interval ($\tau = \frac{1}{20000Hz} = 50\mu s$). The process can be formulated as follows:

$$\mathbf{S}_{x,y}(c) = \begin{cases} 1, & \text{if } \exists t \in ((c-1)\tau, c\tau], V_{x,y}(t) = 0, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

$$V^+_{x,y}(t) = \begin{cases} V^-_{x,y}(t) + I_{x,y}(t), & \text{if } V^-_{x,y}(t) < \Theta, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $V^-_{x,y}(t)$ and $V^+_{x,y}(t)$ denotes the voltage before and after receiving the electric current $I_{x,y}(t)$, $c \in \mathbb{R}$. The spike stream $\mathbf{S}$ is output with the size of $H \times W \times C$ after C times readout during T μs ($C = \frac{T}{\tau}$).

For a blurred image $\mathbf{B}_i$, it can be represented as the average of the clear images $\{\mathbf{L}\}$ over the exposure time, i.e.:

$$\mathbf{B}_i = \frac{1}{T_\mathbf{B}} \int_{T_i - T_\mathbf{B}/2}^{T_i + T_\mathbf{B}/2} \mathbf{L}(t) dt, \tag{3}$$

where T_i is the center of exposure time and $\mathbf{L}_i$ at T_i.

Why spike cameras hold potential on image deblurring. The deblurring process can be formulated as:

$$\arg\min_{\theta_M} \|\hat{\mathbf{L}}_i - \mathbf{L}_i\|, \text{where } \hat{\mathbf{L}}_i = \mathcal{F}(\mathbf{B}_i : \theta_M), \tag{4}$$

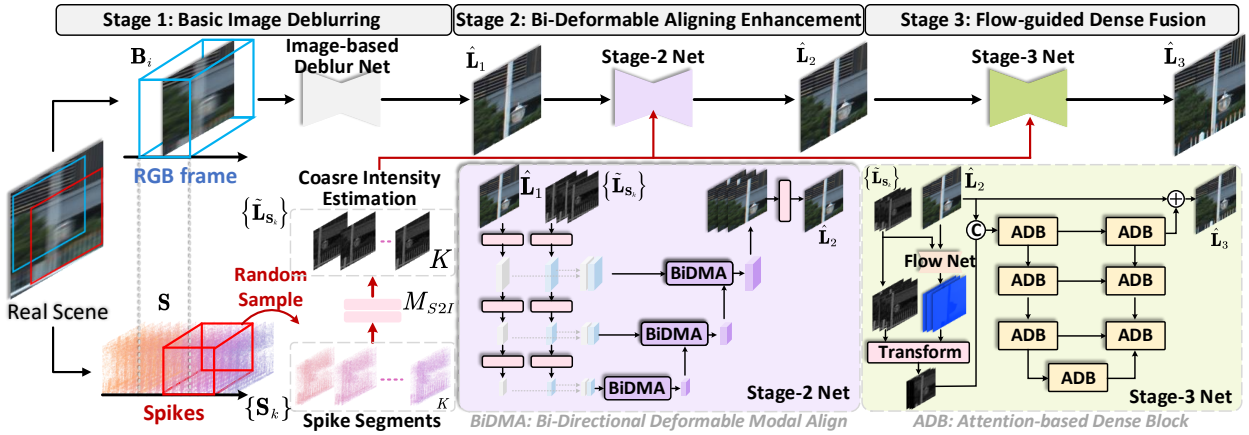

Figure 3. The architecture of the proposed Unaligned Spike-guided Deblur Net (UaSDN).

where θ is the optimized parameters of the model M and $\hat{\mathbf{L}}_i$ is the predicted clear image. The θ is hard to optimize when $\mathbf{B}_i$ holds extreme blur. Neglecting color information and differences in spatial resolution, when the spike camera and RGB camera are perfectly aligned, the clear image $\mathbf{L}_i$ can also be approximated using spikes:

$$\tilde{\mathbf{L}}_i \simeq \mathcal{F}_{M_{S2I}}(\{\mathbf{S}(t)|t \in [T_i - \frac{T_\mathbf{B}}{2}, T_i + \frac{T_\mathbf{B}}{2}]\} : \theta_{S2I}), \tag{5}$$

where M_{S2I} is the model for reconstructing spikes to the approximated image $\tilde{\mathbf{L}}_i$. Many existing models [47, 52] are able to reconstruct high-quality $\tilde{\mathbf{L}}_i$. Despite that $\tilde{\mathbf{L}}_i$ lacks color information and has lower spatial resolution compared to $\mathbf{B}_i$, its clear texture effectively provides deblurring clues for $\mathbf{B}_i$ as shown in Fig. 2, which is proved by our experiments in Sec. 4.5. When not knowing the alignment between the RGB camera and spike camera on both spatial and temporal axis, we cannot get the accurate spikes $\{\mathbf{S}(t)\}$ in Eq. (5). Thus textures in $\tilde{\mathbf{L}}_i$ are not aligned with $\mathbf{L}_i$, which raises challenges for spikes to guide the deblurring.

How to deal with unknown alignment? We set the situation of the unknown alignment as follows: **(1) Spatial axis.** There are blind plane displacements $\Delta\mathbf{P}$ between two cameras. Generally, we set the $\Delta\mathbf{P}$ not a constant but a variable parameter. **(2) Temporal axis.** Spike stream is continuous while RGB frames are discrete. When two streams cannot be precisely synchronized, in real situations, we can roughly fetch a segment of spikes $\mathbf{S}_k$ around the corresponding exposure window T_i of image $\mathbf{B}_i$. The temporal center of $\mathbf{S}_k$ is unknown. In this case, we aim to build the model learning the aligning process adaptively.

3.2. Overall Architecture

For not knowing the alignment, accurately matching the textures between the blurred image and spikes is challenging, requiring the model to have the ability for adaptive feature matching. However, on one hand, the texture of the blurred image $\mathbf{B}$ suffers from significant information loss. On the other hand, the irregular spike data cannot directly express scene textures like images. Therefore, we aim to progressively optimize the motion deblurring, gradually restoring clear textures of the blurred image and spikes. To achieve this, we propose a three-stage model named Unaligned Spike-guided Deblur Net (**UaSDN**). The network architecture of UaSDN is illustrated in Fig. 3.

UaSDN consists of three deblurring stages. *(1) First Stage: Basic Image Deblurring* Unclear texture in blurred images $\mathbf{B}$ makes it difficult to match with features in spikes. We employ a simple network M_1 to learn basic deblurring for $\mathbf{B}$. The network takes the blurred image $\mathbf{B}$ as input and outputs the predicted image $\hat{\mathbf{L}}_1$ *(2) Second Stage: Bi-Deformable Aligning Enhancement.* The texture of $\mathbf{B}$ is enhanced after the stage 1. In this stage, we design a module M_2 based on deformable convolution to achieve basic alignment between spike features and blurred image features. Besides, to restore scene textures in spikes, we use shallow convolutional layers M_{S2I} to input spikes and quickly infer the rough light intensity. We randomly sample K segments of spikes $\{\mathbf{S}_k|k \in [1, 2, ..K]\}$ as guidance input to M_2, which outputs the enhanced image $\hat{\mathbf{L}}_2$. *(3) Third Stage: Flow-guided Dense Fusion.* In this stage, we aim to utilize optical flows v_{S2I} from spikes $\mathbf{S}$ to the image $\hat{\mathbf{L}}_2$ as guidance for precise alignment. To achieve this, we design the multi-scale flow-guided deblurring module, which takes the sampled K spike segments as input and predicts the final clear image $\hat{\mathbf{L}}_3$. The process can be formulated as:

$$\hat{\mathbf{L}}_1 = \mathcal{F}_{M_1}(\mathbf{B} : \theta_{M_1}), \tag{6}$$

$$\hat{\mathbf{L}}_2 = \mathcal{F}_{M_2}(\hat{\mathbf{L}}_1, \{\mathbf{S}_k|k \in [1, 2, \ldots, K]\} : \theta_{M_2}), \tag{7}$$

$$\hat{\mathbf{L}}_3 = \mathcal{F}_{M_3}(\hat{\mathbf{L}}_2, \{\mathbf{S}_k|k \in [1, 2, \ldots, K]\} : \theta_{M_3}). \tag{8}$$

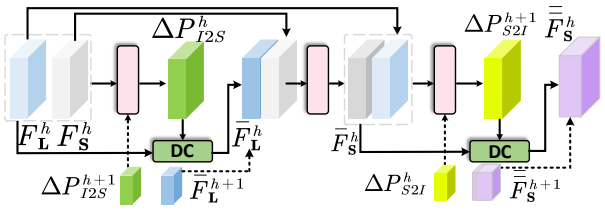

Figure 4. The architecture of the proposed Bi-directional Deformable Modal Aligning (BiDMA) module.

3.3. Stage 1: Basic Image Deblurring

In the first stage, we employed one of the current state-of-the-art image restoration networks, NAFNet [5], as the basic deblurring module M_1. We aim to use M_1 to lighten the blurring effect and enhance the texture of the blurred image $\mathbf{B}$, as shown in Fig. 3. Specifically, NAFNet is a U-shaped multi-scale network. $\mathbf{B}$ undergoes 4 times down-sampling during feature extraction, followed by up-sampling. In the up-sampling process, features from each scale are fused, and the network finally outputs $\hat{\mathbf{L}}_1$. To reduce parameter count and accelerate inference speed, the overall network channel number is halved. Detailed architectures are provided in the supplementary material.

3.4. Stage 2: Bi-Deformable Aligning Enhancement

At this stage, spikes are introduced. As in Eq. (5), if two cameras are perfectly aligned in space and time, spikes can be optimized to obtain a clear grayscale image $\tilde{\mathbf{L}}_\mathbf{S}$ without color information and at a lower resolution. The reconstructed gray-scale image can effectively aid in deblurring the RGB image. Our goal is to learn an adaptive pixel alignment relationship between modalities through learning. We first pass the sampled spike fragments through a shallow convolutional module M_{S2I} to quickly estimate the corresponding rough grayscale intensity $\tilde{\mathbf{L}}_\mathbf{S}$. M_{S2I} consists of only 4 convolutional layers, which is formulated as:

$$\tilde{\mathbf{L}}_\mathbf{S} = \mathcal{F}_{M_{S2I}}(\mathbf{S} : \theta_{M_{S2I}}), \tag{9}$$

In addition, we designed the bi-directional deformable modal align (BiDMA) module, as illustrated in Fig. 4. The spike and image features are extracted with downsampling, denoted as $\mathbf{F_S} = \{F_\mathbf{S}^h\}$ and $\mathbf{F_L} = \{F_\mathbf{L}^h\}$, where $h \in [1, 2, ..., H]$ is the number of scales. A deformable convolution network(DCN) [10] $\mathcal{F}_{DCN}$ can align two kinds of features, the principle can be denoted as:

$$F(p) = \sum_{q=1}^{Q} w_q \cdot F(p + p_q + \Delta p_q) \cdot \Delta m_q, \tag{10}$$

where $p = (x, y)$ is the coordinate of the center pixel, Q denotes pixel numbers in the neighbor of p, p_q is the fixed offset from p, w_q denotes the weight, Δm_q denotes the modulation scalar. Δp_q is the learnable offset, which is predicted by two modal features with a module $\mathcal{F}_\Delta$ consisting of several convolution layers. In each h-level, we consider the bi-directional deformable aligning: 1) align $F_\mathbf{L}^h$ of image to $F_\mathbf{S}^h$ of spikes and get the $\bar{F}_\mathbf{L}^h$; 2) fuse the $\bar{F}_\mathbf{L}^h$ and $F_\mathbf{S}^h$, and get $\bar{F}_\mathbf{S}^h$; 3) align the $\bar{F}_\mathbf{S}^h$ back to $F_\mathbf{L}^h$, and get $\bar{\bar{F}}_\mathbf{S}^h$. The process can be formulated as:

$$\begin{aligned} \Delta P_{I2S}^h &= \mathcal{F}_\Delta([F_\mathbf{L}^h, F_\mathbf{S}^h], \Delta P_{I2S}^{h+1}), \\ \bar{F}_\mathbf{L}^h &= \mathcal{F}_\oplus(\mathcal{F}_{DCN}(F_\mathbf{L}^h, \Delta P_{I2S}^h), \mathbf{Up}(\bar{F}_\mathbf{L}^{h+1})), \\ \bar{F}_\mathbf{S}^h &= \mathcal{F}_\oplus(\bar{F}_\mathbf{L}^h, F_\mathbf{S}^h), \end{aligned} \tag{11}$$

$$\begin{aligned} \Delta P_{S2I}^h &= \mathcal{F}_\Delta([\bar{F}_\mathbf{S}^h, F_\mathbf{L}^h], \Delta P_{S2I}^{h+1}), \\ \bar{\bar{F}}_\mathbf{S}^h &= \mathcal{F}_\oplus(\mathcal{F}_{DCN}(\bar{F}_\mathbf{S}^h, \Delta P_{S2I}^h), \mathbf{Up}(\bar{\bar{F}}_\mathbf{S}^{h+1})), \\ F_{align} &= \mathcal{F}_\oplus(\bar{\bar{F}}_\mathbf{S}^1, F_\mathbf{L}^1), \end{aligned} \tag{12}$$

where $\mathcal{F}_{DCN}$ is the DCN network as in Eq. (10), $\mathcal{F}_\oplus$ is the feature fusion function with simple Conv layers, $\mathbf{Up}(\cdot)$ is the upsampling operation. Eq. (11) describe the align process from images to spikes, and Eq. (12) describe the align process from spikes back to images. In this way, at the 1-level, the aligned feature F_{align} is fused by $\mathcal{F}_\oplus$ with several Conv layers. As described in Sec. 3.1, we sample K segments of spikes for alignment with the image. Thus the output $\hat{\mathbf{L}}_2$ of stage 2 is formulated as:

$$\hat{\mathbf{L}}_2 = \mathcal{F}_{pred}([F_{align,1}, F_{align,2}, ..., F_{align,K}]). \tag{13}$$

3.5. Stage 3: Flow-guided Dense Fusion

After the alignment in stage 2, textures of $\hat{\mathbf{L}}_2$ are further enhanced by spikes. In this stage, we aim to achieve precise alignment of spike features through optical flow guidance, improving the deblurring effect by supplementing detailed textures from features in the spike. Thus, we build a flow-guided dense fusion network for the final refinement. as illustrated in Fig. 3. Firstly, K coarse light intensity estimation $\{\tilde{\mathbf{L}}_{\mathbf{S}_k} | k \in [1, 2, \ldots, K]\}$ are predicted as Eq. (9) described from the sampled spike segments $\{\mathbf{S}_k | k \in [1, 2, \ldots, K]\}$. Then, we compute the optical flows $V_{\mathbf{S}_k \to \hat{\mathbf{L}}_2}$ between the $\hat{\mathbf{L}}_2$ and each $\tilde{\mathbf{L}}_{\mathbf{S}_k}$, then warp the $\tilde{\mathbf{L}}_{\mathbf{S}_k}$ to $\tilde{\mathbf{L}}_{\mathbf{S}_k}^w$ according to the $V_{\mathbf{S}_k \to \hat{\mathbf{L}}_2}$. The process is formulated as:

$$\begin{aligned} V_{\mathbf{S}_k \to \hat{\mathbf{L}}_2} &= \mathcal{F}_{flow}(\tilde{\mathbf{L}}_{\mathbf{S}_k}, \hat{\mathbf{L}}_2), \\ \tilde{\mathbf{L}}_{\mathbf{S}_k}^w &= \mathcal{G}(\tilde{\mathbf{L}}_{\mathbf{S}_k}, V_{\mathbf{S}_k \to \hat{\mathbf{L}}_2}), \end{aligned} \tag{14}$$

where $\mathcal{F}_{flow}$ is the pretrained image-based flow estimation network [35], and $\mathcal{G}$ denotes the feature warp operator. Then, we concatenate $\{\tilde{\mathbf{L}}_{\mathbf{S}_k}^w | k \in [1, 2, \ldots, K]\}$ and $\hat{\mathbf{L}}_2$, and get the tiled image $\hat{\mathbf{L}}_{tile}$. The U-shape network contains the encoder and decoder. In the encoder, the image features of $\hat{\mathbf{L}}_{tile}$ are extracted with downsampling, denoted as $\mathbf{F}_{\hat{\mathbf{L}}_{tile}} = \{F_{\hat{\mathbf{L}}_{tile}}^h\}$, $h = [1, 2, ..., H]$. The feature $F_{\hat{\mathbf{L}}_{tile}}^h$ at

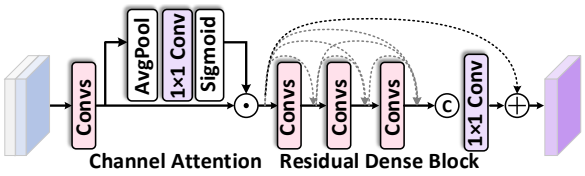

Figure 5. The proposed Attention-base Dense Block (ADB).

level h doubles the channel number of $F^{h-1}_{\hat{\mathbf{L}}_{tile}}$ at level $h-1$. In the decoder, features from the previous layer are upsampled through convolution layers and a pixel shuffle operation, then fuse with features at the current level. Each encoder and decoder at each level is an attention-base dense block(ADB) consisting of a channel attention layer (CA) and a residual dense block (RDB) [50], as in Fig. 5. The CA fuses two features from two modalities and the RDB aims to extract features effectively. In this way, the final deblurred image $\hat{\mathbf{L}}_3$ is predicted with the sum of the output of several convolutional layers and the input $\hat{\mathbf{L}}_2$.

3.6. Loss functions

We train the three stages separately, for each stage, we use the $\mathcal{L}_1$ loss to optimize the network. For the coarse light intensity estimation module M_{S2I}, $\mathcal{L}_1$ loss is adopted too.

$$\begin{aligned} \mathcal{L}^i_{deblur} &= |\hat{\mathbf{L}}^i_1 - \mathbf{L}_{rgb}|, \text{where } i = 1, 2, 3, \\ \mathcal{L}^i_{M_{S2I}} &= |\tilde{\mathbf{L}}_{\mathbf{S}} - \mathbf{L}_{gray}|, \end{aligned} \tag{15}$$

where i is the stage index, $\mathbf{L}_{rgb}$ is the ground-truth clear RGB image, and $\mathbf{L}_{gray}$ is the clear gray image with equal spatial resolution as spikes.

4. Experiment

4.1. Datasets and Training Settings

To train the network, we utilize datasets containing high-speed RGB images as the base and simulate the principle of a spike camera to generate continuous spike streams. Specifically, we conduct training and testing on two large datasets: the X4k1000FPS [30] and the REDS dataset [25]. The X4K1000FPS dataset includes high-definition 1000fps videos, while REDS comprises 1000fps videos, along with real or synthesized blurred images. To generate spike data, we employ the state-of-the-art video interpolation algorithm EMA-VFI [46] to increase the frame rate by 8 times for the image sequences in both datasets, resulting in ultra-high frame rate videos. Ultra-high frame rate videos can approximate the continuous changes in scene lighting, allowing us to use the working principle of a spike camera to generate corresponding spike streams as shown in Eq. (1). Additionally, since the X4K1000FPS dataset does not include blurred images, we simulate the blurry images using an average of 33 frames. In X4K1000FPS and REDS, the equivalent exposure time for the blurred images is 1/120 and 1/24 seconds, respectively.

Method	Input Data	PSNR ↑	SSIM ↑
HINet [4]	Image	33.74	0.935
NAFNet [5]	Image	34.13	0.937
EFNet [32]	Image+Spike	33.28	0.928
REFID [33]	Image+Spike	34.22	0.939
SpkDeblurNet [7]	Image+Spike	34.47	0.941
UaSDN(Ours)	Image+Spike	**35.78**	**0.964**

Table 1. Comparison of various motion deblurring methods on X4K1000FPS [30].

Method	Input Data	PSNR ↑	SSIM ↑
HINet [4]	Image	31.10	0.902
NAFNet [5]	Image	30.73	0.894
EFNet [32]	Image+Spike	31.74	0.912
REFID [33]	Image+Spike	31.23	0.904
SpkDeblurNet [7]	Image+Spike	31.16	0.911
UaSDN(Ours)	Image+Spike	**33.61**	**0.942**

Table 2. Comparison of various motion deblurring methods on REDS [25].

How do we set the unknown spatiotemporal alignment between RGB frames and spikes? In the training set, along the spatial axis, we introduce random offsets in *up, down, left, and right* directions for the spikes compared to the blurred image, with a maximum offset of **10%**. Along the temporal axis, the center timestamp of the sampled spike segments compared to the blurred image was set with random offsets in *forward and backward* directions, with a maximum offset of **25%**. For testing, we set the above offset scheme as a fixed uniform distribution to ensure that all methods use the same test data in evaluation. Details of the dataset and training are in supplementary materials.

4.2. Results on X4K1000FPS

To demonstrate the deblurring capability of our proposed model **UaSDN**, under modality misalignment, we compare it with five methods: two image-based networks, HINet [4] and NAFNet [5], two state-of-the-art event-based networks, EFNet [32] and REFID [33], and the spike-based SpkDeblurNet [7]. To ensure fairness in the experiments, we modify the inputs of EFNet and REFID to be the blurred images and spikes, with the spike length matching that of UaSDN. We transform the spikes to the input representation used by EFNet and REFID.

As in Tab. 1, our method achieves 35.78dB in PSNR and 0.964 in SSIM, demonstrating a significant performance improvement compared to other methods. Compared to methods that only take the image as input, our approach

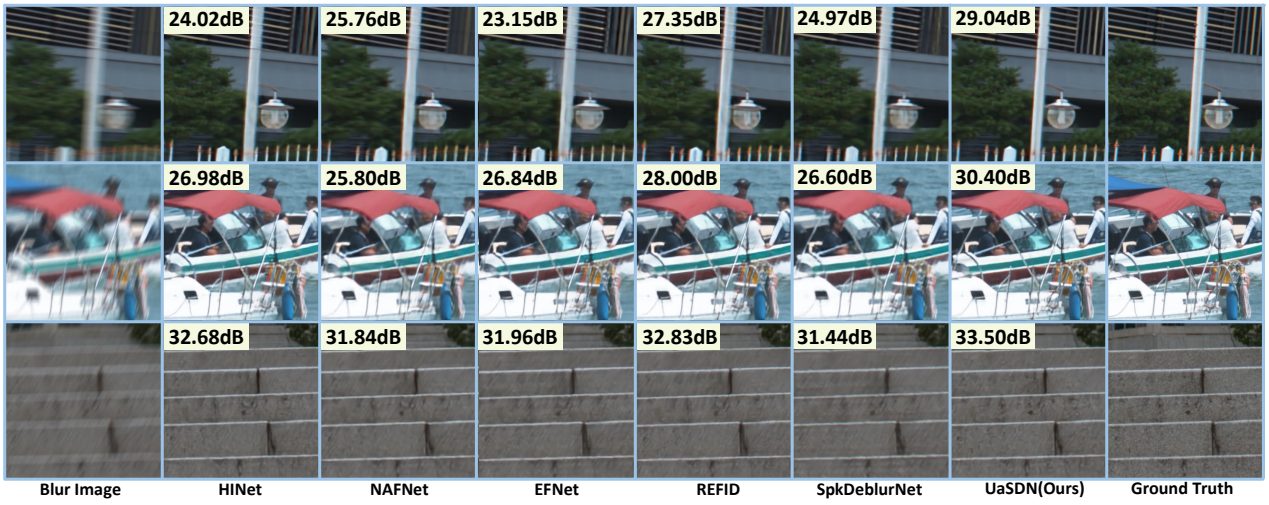

Figure 6. Visualized results of our method (UaSDN) on X4K1000FPS compared with HINet, NAFNet, EFNet, REFID and SpkDeblurNet.

Figure 7. Visualized results of our method (UaSDN) on REDS dataset compared with HINet, NAFNet, EFNet, REFID and SpkDeblurNet.

outperforms HINet and NAFNet by 2.04dB and 1.65dB in PSNR, respectively. It indicates the effectiveness of utilizing spikes as guidance for deblurring. Furthermore, compared to the other two event-based networks that take spikes as input, our approach surpasses EFNet and REFID by 2.50dB and 1.56dB in PSNR. It strongly validates the effectiveness of our model in aligning the two modal data in both time and space. As in Fig. 6, the visual results demonstrate that our method effectively obtains fine textures. For example, in the first row, our method successfully enhances the blurred textures without introducing any shape distortion, while other methods fail to address such extreme blurriness and result in shape distortions. It proves the precise spatiotemporal feature alignment achieved by our method.

4.3. Results on REDS

On the REDS dataset, as shown in Tab. 2, our method achieves a PSNR of 33.61dB and an SSIM of 0.942. Compared to the image-based methods HINet and NAFNet, our method surpasses them by 2.51dB and 2.88dB, respectively. Compared to the event-based methods EFNet and REFID, our method outperforms them by 1.87dB and 2.38dB. As illustrated in Fig. 7, our method can significantly enhance deblurring performance in the presence of complex textures, such as text on license plates, text on ships, and details in faces. Through the comparison of the REDS dataset, we further demonstrate the deblurring capability of the model and its generalization ability to different datasets.

4.4. Results on Real World Spikes

To assess the deblurring capability of our method in real-world scenarios, we set up a simple unaligned spike-RGB hybrid camera system. Fig. 8 shows real-world test cases, including fast-moving objects at close range and outdoor driving scenarios with rapid motion. As shown in Fig. 8, our method significantly outperforms other approaches in

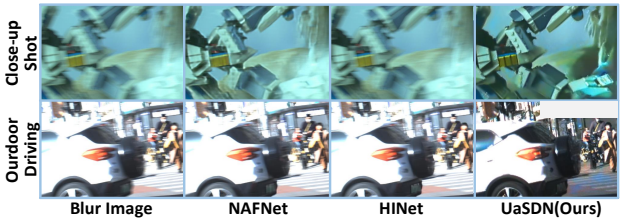

Figure 8. Visualized results on real-world data captured with an RGB camera and a spike camera.

Stage1	Stage2	Stage3	PSNR ↑	SSIM ↑
✓	×	×	32.30	0.914
✓	✓	×	32.90	0.923
✓	✓	✓	**35.78**	**0.964**

Table 3. Experiments on the performance of the model stage.

motion deblurring. We also observed that due to the high dynamic information contained in the spike stream, our method excels of deblurring in scenes with overexposure.

4.5. Ablation Studies

A. Performance Increases at each stage. To demonstrate the effectiveness of multi-stage model design, we conduct ablation experiments on the deblurring effects at each stage. As shown in Tab. 3, when performing the first stage training, the model possesses basic deblurring capabilities. With the introduction of the bi-directional deformable alignment in the second stage, the model shows a noticeable improvement of 0.6dB and 0.01 in PSNR and SSIM, respectively. The addition of the flow-guided feature fusion network in the third stage further increases the PSNR to the highest value of 35.78 dB. These experimental results indicate that through progressive training, clear texture information from the spike stream can effectively be introduced into the image domain, guiding the deblurring process.

B. Deblurring Potentials under Precise Alignment. The above experiments demonstrate that our method effectively removes blur when spikes and images are unknownly aligned in both time and space. To further validate the potential of our method, we modify the X4k1000FPS and REDS datasets to strictly align spikes and images and test the model's performance under the condition of complete spatiotemporal alignment between modalities. As shown in Tab. 4, the results on the X4K1000FPS dataset include the quantitative results of two other methods, EFNET and REFID, which also take spikes as guiding input. The results show that our method achieves a PSNR of 36.92dB, surpassing the other two methods by 0.56dB and 0.62dB. We also test the performance improvement on REDS. The experiment shows that our method improves from 33.61dB to 34.60dB on REDS, indicating that our method has great potential for application in spike-guided deblurring. Visualized results are shown in the supplementary materials.

Method	Input Data	PSNR ↑	SSIM ↑
EFNet [32]	Image+Spike	36.36	0.960
REFID [33]	Image+Spike	36.30	0.962
UaSDN(Ours)	Image+Spike	**36.92**	**0.969**

Table 4. Comparison of various motion deblurring methods on X4K1000FPS [30] when spikes and images are precise aligned.

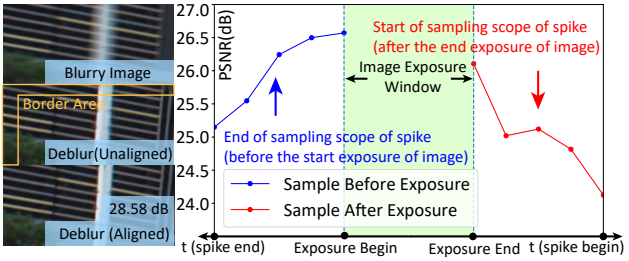

Figure 9. Influence of border areas when sampled spikes are totally not involved in the exposure of the image.

C. How the border areas are handled from spatial and temporal perspectives. Generally, the motion near the border is similar. The model learns motion clues in spikes that supplement textures. Besides, our image-based stage-1 in UaSDN performs basic deblurring without spikes. In Fig. 9, the image that is spatiotemporal unaligned with spikes is recovered well by UaSDN. The quality is close to the deblurred one with perfect alignment (28.58 dB). The curve shows the model maintains acceptable results though the PSNR declines when spikes are not involved in the image exposure time at all.

5. Conclusion

We first explore the motion deblurring guided by a spike camera under unknown spatiotemporal alignment between two modalities. We believe the proposed method possesses potential in the real world. The proposed three-stage UaSDN contains a bi-directional deformable alignment and a flow-based fusion module. Results prove that UaSDN outperforms other methods on unaligned deblurring.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (62176003, 62088102, 62306015), the China Postdoctoral Science Foundation (2022M720238, 2023T160015), the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001), and the Beijing Nova Program (20230484362).

References

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

[2] Mingdeng Cao, Zhihang Zhong, Yanbo Fan, Jiahao Wang, Yong Zhang, Jue Wang, Yujiu Yang, and Yinqiang Zheng. Towards real-world video deblurring by exploring blur formation process. In *European Conference on Computer Vision*, pages 327–343. Springer, 2022. 2

[3] Yakun Chang, Chu Zhou, Yuchen Hong, Liwen Hu, Chao Xu, Tiejun Huang, and Boxin Shi. 1000 fps hdr video with a spike-rgb hybrid camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22180–22190, 2023. 2

[4] 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. 1, 2, 3, 6

[5] 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. 1, 3, 5, 6

[6] Shiyan Chen, Chaoteng Duan, Zhaofei Yu, Ruiqin Xiong, and Tiejun Huang. Self-supervised mutual learning for dynamic scene reconstruction of spiking camera. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 2859–2866. International Joint Conferences on Artificial Intelligence Organization, 2022. 2

[7] Shiyan Chen, Jiyuan Zhang, Yajing Zheng, Tiejun Huang, and Zhaofei Yu. Enhancing motion deblurring in high-speed scenes with spike streams. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 6

[8] Hoonhee Cho, Yuhwan Jeong, Taewoo Kim, and Kuk-Jin Yoon. Non-coaxial event-guided motion deblurring with spatial alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12492–12503, 2023. 3

[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. 2

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–773, 2017. 5

[11] Peiqi Duan, Yi Ma, Xinyu Zhou, Xinyu Shi, Zihao W Wang, Tiejun Huang, and Boxin Shi. Neurozoom: Denoising and super resolving neuromorphic events and spikes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[12] Tiejun Huang, Yajing Zheng, Zhaofei Yu, Rui Chen, Yuan Li, Ruiqin Xiong, Lei Ma, Junwei Zhao, Siwei Dong, Lin Zhu, et al. 1000× faster camera and machine vision with ordinary devices. *Engineering*, 2022. 1, 2

[13] Bangrui Jiang, Zhihuai Xie, Zhen Xia, Songnan Li, and Shan Liu. Erdn: Equivalent receptive field deformable network for video deblurring. In *European Conference on Computer Vision*, pages 663–678. Springer, 2022. 2

[14] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning event-based motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3320–3329, 2020. 3

[15] Taewoo Kim, Jeongmin Lee, Lin Wang, and Kuk-Jin Yoon. Event-guided deblurring of unknown exposure time videos. In *European Conference on Computer Vision*, pages 519–538. Springer, 2022. 3

[16] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5886–5895, 2023. 1, 2

[17] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8183–8192, 2018. 2

[18] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8878–8887, 2019. 2

[19] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 3

[20] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 2024. 3

[21] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15μ s latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43 (2):566–576, 2008. 1

[22] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning event-driven video deblurring and interpolation. In *European Conference on Computer Vision*, pages 695–710. Springer, 2020. 3

[23] Diederik Paul Moeys, Federico Corradi, Chenghan Li, Simeon A Bamford, Luca Longinotti, Fabian F Voigt, Stewart Berry, Gemma Taverni, Fritjof Helmchen, and Tobi Delbruck. A sensitive dynamic and active pixel vision sensor for color or neural imaging applications. *IEEE Transactions on Biomedical Circuits and Systems*, 12(1):123–136, 2017. 1

[24] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017. 2

[25] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu

Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 6

[26] Jihyong Oh and Munchurl Kim. Demfi: deep joint deblurring and multi-frame interpolation with flow-guided attentive correlation and recursive boosting. In *European Conference on Computer Vision*, pages 198–215. Springer, 2022. 2

[27] Jinshan Pan, Boming Xu, Jiangxin Dong, Jianjun Ge, and Jinhui Tang. Deep discriminative spatial and temporal network for efficient video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22191–22200, 2023. 2

[28] 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. 1, 3

[29] Wei Shang, Dongwei Ren, Dongqing Zou, Jimmy S Ren, Ping Luo, and Wangmeng Zuo. Bringing events into video deblurring with non-consecutively blurry frames. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4531–4540, 2021. 3

[30] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. Xvfi: extreme video frame interpolation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14489–14498, 2021. 6, 8

[31] Maitreya Suin and AN Rajagopalan. Gated spatio-temporal attention-guided video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7802–7811, 2021. 2

[32] 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. 1, 2, 3, 6, 8

[33] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18043–18052, 2023. 2, 3, 6, 8

[34] 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. 2

[35] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 5

[36] Minggui Teng, Chu Zhou, Hanyue Lou, and Boxin Shi. Nest: Neural event stack for event-based image enhancement. In *European Conference on Computer Vision*, pages 660–676. Springer, 2022. 3

[37] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *European Conference on Computer Vision*, pages 155–171. Springer, 2020. 3

[38] Yusheng Wang, Yunfan Lu, Ye Gao, Lin Wang, Zhihang Zhong, Yinqiang Zheng, and Atsushi Yamashita. Efficient video deblurring guided by motion magnitude. In *European Conference on Computer Vision*, pages 413–429. Springer, 2022. 2

[39] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 3

[40] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based blurry frame interpolation under blind exposure. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1588–1598, 2023. 3

[41] Lujie Xia, Jing Zhao, Ruiqin Xiong, and Tiejun Huang. Svfi: spiking-based video frame interpolation for high-speed motion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2910–2918, 2023. 2

[42] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblurring with real events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2583–2592, 2021. 3

[43] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *European Conference on Computer Vision*, pages 492–511. Springer, 2020. 3

[44] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14821–14831, 2021. 2

[45] 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 International Conference on Computer Vision*, 2022. 3

[46] Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5682–5692, 2023. 6

[47] Jiyuan Zhang, Shanshan Jia, Zhaofei Yu, and Tiejun Huang. Learning temporal-ordered representation for spike streams based on discrete wavelet transforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 137–147, 2023. 2, 4

[48] 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. 3

[49] 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. 3

[50] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2480–2495, 2020. 6

[51] Jing Zhao, Jiyu Xie, Ruiqin Xiong, Jian Zhang, Zhaofei Yu, and Tiejun Huang. Super resolve dynamic scene from continuous spike streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2533–2542, 2021. 2

[52] Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, and Tiejun Huang. Spk2imgnet: Learning to reconstruct dynamic scene from continuous spike stream. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11996–12005, 2021. 2, 4

[53] Yajing Zheng, Lingxiao Zheng, Zhaofei Yu, Boxin Shi, Yonghong Tian, and Tiejun Huang. High-speed image reconstruction through short-term plasticity for spiking cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6358–6367, 2021. 2

[54] Yajing Zheng, Zhaofei Yu, Song Wang, and Tiejun Huang. Spike-based motion estimation for object tracking through bio-inspired unsupervised learning. *IEEE Transactions on Image Processing*, 32:335–349, 2022. 2

[55] Yajing Zheng, Lingxiao Zheng, Zhaofei Yu, Tiejun Huang, and Song Wang. Capture the moment: High-speed imaging with spiking cameras through short-term plasticity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):81274–8142, 2023. 2

[56] Zhihang Zhong, Mingdeng Cao, Xiang Ji, Yinqiang Zheng, and Imari Sato. Blur interpolation transformer for real-world motion from blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5713–5723, 2023. 2

[57] Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. A retina-inspired sampling method for visual texture reconstruction. In *IEEE International Conference on Multimedia and Expo*, pages 1432–1437. IEEE, 2019. 2

[58] Qi Zhu, Man Zhou, Naishan Zheng, Chongyi Li, Jie Huang, and Feng Zhao. Exploring temporal frequency spectrum in deep video deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12428–12437, 2023. 2