Hybrid All-in-focus Imaging from Neuromorphic Focal Stack

Minggui Teng, Hanyue Lou, Yixin Yang, Tiejun Huang, *Senior Member, IEEE*, and Boxin Shi, *Senior Member, IEEE*

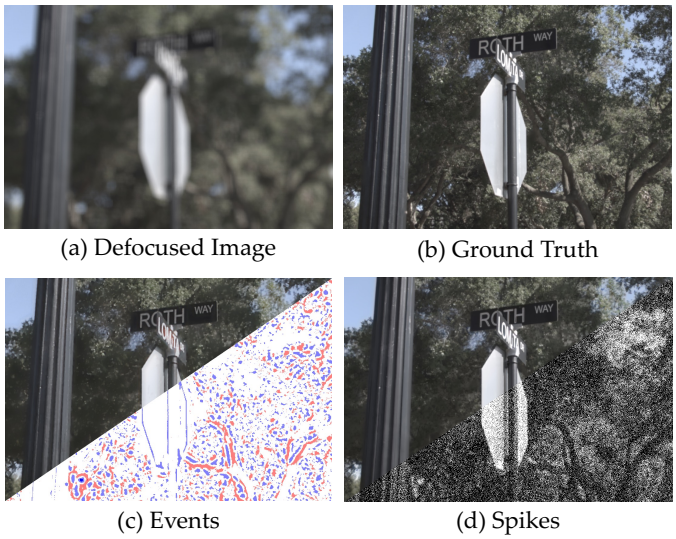
Abstract—Creating an image focal stack requires multiple shots, which captures images at different depths within the same scene. Such methods are not suitable for scenes undergoing continuous changes. Achieving an all-in-focus image from a single shot poses significant challenges, due to the highly ill-posed nature of rectifying defocus and deblurring from a single image. In this paper, to restore an all-in-focus image, we introduce the neuromorphic focal stack, which is defined as neuromorphic signal streams captured by an event/ a spike camera during a continuous focal sweep, aiming to restore an all-in-focus image. Given an RGB image focused at any distance, we harness the high temporal resolution of neuromorphic signal streams. From neuromorphic signal streams, we automatically select refocusing timestamps and reconstruct corresponding refocused images to form a focal stack. Guided by the neuromorphic signal around the selected timestamps, we can merge the focal stack using proper weights and restore a sharp all-in-focus image. We test our method on two distinct neuromorphic cameras. Experimental results from both synthetic and real datasets demonstrate a marked improvement over existing state-of-the-art methods.

Index Terms—Neuromorphic Camera, All-in-focus Imaging, Hybrid Camera System.

1 Introduction

The brightness of the image depends on how much light the camera lens lets in through its aperture. A larger aperture setting keeps a good balance between signal and noise even with shorter exposure time. This is crucial for taking pictures of fast-moving subjects or in low light scenarios while reducing noise. However, larger aperture also results in a shallow depth of field (DoF), causing parts of the image to blur. This selective focus can be artistically leveraged in instances like portrait photography, highlighting the main subject against a blurry background. Yet, for purposes like detailed microscopy imaging [30], it is often desired to have the entire image in focus. An all-in-focus image also benefits high-end vision tasks, including object detection [35] and semantic segmentation [13].

Achieving an all-in-focus image by deblurring a defocused one is challenging. This is because the defocus effect, influenced by the aperture shape and scene depth, often changes across the image and is hard to be estimated precisely [60]. Traditional methods [11], [17], [44] often work in two stages. They first estimate the pixel-wise or patch-wise defocus kernels with image priors and then apply non-blind image deconvolution to each pixel or patch. A paradigm shift is noted with contemporary end-to-end deep learning methodologies [22], [38], [39], [47]. Powered by large sets of training data, these methods learn from paired defocused and sharp images and generally outperform the older two-step methods. However, they have their limitations. Sometimes, they still produce images with unwanted visual effects, like ringing artifacts or lingering blur in high-frequency zones, particularly when faced with regions that are both weakly textured and heavily defocused.



(a) Defocused Image (b) Ground Truth (c) Events (d) Spikes

Figure 1: An example result of all-in-focus imaging guided by neuromorphic signals. From a defocused image (a) and its corresponding neuromorphic focal stack, our approach produces an all-in-focus image that is aided with a neuromorphic focal stack formed with events (c) or spikes (d), closely matching the clarity of the ground truth (b). In both (c) and (d), the upper-left section shows the reconstructed all-in-focus image, while the lower-right depicts the corresponding event/spike frame.

To overcome the ill-posedness of estimating the defocus kernel from a single image, merging a focal stack, *i.e.*,

- *Minggui Teng, Hanyue Lou, Yixin Yang, Tiejun Huang, and Boxin Shi are with the State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing, 100871, China.*
- *E-mails: {minggui_teng,hylz,yangyixin93,tjhuang,shiboxin}@pku.edu.cn.*
- *Corresponding author: Boxin Shi.*

a sequence of images taken at different focus distances, can generate an all-in-focus image reliably [14], [49], [59]. However, capturing a focal stack requires a static scene and multiple exposures. Moreover, the selection of focus distances is a key factor in capturing the focal stack, which requires elaborate design.

Neuromorphic cameras, cutting-edge sensor innovations, fall into two primary categories: differential-based [6], [43] and integral-based [16], distinguished by their signal correlation with scene radiance. Differential-based, well-known as *event cameras* [6], [43], detect brightness fluctuations and initiate an event when the log difference surpasses a predetermined threshold. Integral-based, well-known as *spike cameras* [16], on the other hand, allow pixels to be continuously exposed, triggering a spike signal once the intensity accumulation crosses a preset threshold. Despite their distinct formulations, both types own remarkable temporal resolution, capturing near-continuous intensity variations of scenes. This enables the generation of high-speed videos from neuromorphic signal streams [34], [50]–[52], [55], [61], prompting us to contemplate: *Can neuromorphic streams, forming "focal stacks", aid all-in-focus imaging?*

In our prior work [29], we proposed the *event focal stack* (EFS) approach to achieve all-in-focus imaging for the first time. By implementing a focal sweep with the event camera, we derived the EFS. Benefiting from the high-temporal information available, we further enhanced the defocused RGB image to broaden its DoF. However, an event signal captures only intensity changes, leading to a lack of texture information in areas with noticeable changes. As mentioned above, another type of neuromorphic camera, the spike camera, allows continuous exposure for each pixel. This characteristic yields richer texture observations during the focal sweep, as depicted in Figure 1 (the lower-right sections of (c) and (d) underscore the superior texture detailing of spike signals compared to event signals). Information enhancement solutions for event signals have been successfully adapted to spike signals in a unified framework [9]. This revelation underscores the need to modify our original method [29] to accommodate both neuromorphic camera variants. The previous focal stack merging network tailored to event signals [29] also has a limitation which demands a fixed number of image inputs and subsequently caps the size of the reconstructed image focal stack. The quantity of input images also impacts the quality of the all-in-focus image restoration.

In this paper, we introduce the *neuromorphic focal stack* (NFS) concept for all-in-focus imaging, which operates on both event and spike cameras. It is composed of neuromorphic signal (event/spike) streams obtained from a continuous focal sweep with a neuromorphic camera, which can be used to reconstruct an image focal stack (given an RGB image focused at an arbitrary distance) and predict the merging weights for all-in-focus image recovery, as shown in Figure 1. This paper extends [29] and makes the following contributions:

- We reformulate the all-in-focus imaging pipeline to incorporate neuromorphic cameras, covering both event and spike cameras from the perspective of image focal stack merging. Building upon this pipeline, we introduce a non-learning-based recovery method by firmly establishing a connection between defocused and all-in-focus images using neuromorphic signals.
- We improve the previous EvMergeNet which merged image focal stacks with a fixed number of inputs, to the updated NeuroMergeNet, which can manage arbitrary image inputs by forecasting initial weights for each image, and subsequently converting these preliminary weights into final merging weights.
- We propose a unified data-driven framework designed to capitalize high-speed temporal information within neuromorphic signals (either spike or event) for mitigating RGB image defocus.

We quantitatively and qualitatively evaluate our method on both synthetic and real datasets and demonstrate its superior quality in recovering all-in-focus images over state-of-the-art methods.

2 Related Work

In this section, we briefly review all-in-focus image recovery methods in two categories: image-based methods and computational photography methods. The inputs for image-based methods are obtained using conventional cameras with a single shot, while computational photography methods use a specific capture pipeline or unconventional lenses or sensors. The neuromorphic-based video reconstruction methods, which are partially related to image focal stack generation from neuromorphic signals, are also reviewed.

Image-based methods. Conventional defocus deblurring methods [11], [17], [44] usually contain two steps: estimating the defocus map and applying non-blind deconvolution for deblurring. The quality of deblurred results highly depends on the accuracy of the defocus map. To boost the performance of defocus map estimation, Park *et al*. [33] fused multi-scale image features and hand-crafted features to improve the accuracy of the defocus map. Lee *et al*. [21] proposed a domain adaptation method to transfer features of a synthetic defocused image to the real blurred one for reconstructing a more realistic defocus map. Zhao *et al*. [57] proposed an adversarial promoting learning framework to estimate defocus maps in a weakly-supervised manner.

To avoid the reliance on defocus map estimation in two-step approaches, recent end-to-end defocus deblurring networks have demonstrated higher robustness and performance. Lee *et al*. [22] proposed an Iterative Filter Adaptive Network (IFAN) to handle spatially-varying and large defocus blur via predicting filters for defocused features. Son *et al*. [47] proposed a Kernel-sharing Parallel Atrous Convolutional (KPAC) block to handle defocus blur with slightly varying shapes, which simulates the varying scales of inverse kernels. Ruan *et al*. [38] proposed a neural network trained on both light field generated and real defocused images to enhance the defocus deblurring performance. However, it is hard to recover the high-frequency regions from the defocused image, and the artifacts become obvious when applying deconvolution on a single defocus image. Thus, it is desirable to obtain all-in-focus images using a

more robust method, which can record the continuous scene and depth information.

Computational photography methods. Computational photography based defocus deblurring methods utilize specific capture pipelines (*e.g.*, focal stack [14], [59], focal sweep [20]) or unconventional lens (*e.g.*, coded aperture [23], wavefront coding [8], lattice lens [24]) to relieve the ill-posedness of defocus deblurring. Recently, Abuolaim *et al*. [1], [2] have illustrated that the difference between the two views of a dual-pixel image is related to the defocus amount and can be utilized to further improve the defocus deblurring performance. Although additionally useful cues for all-in-focus image recovery (than single image-based methods) have been encoded and decoded via various computational photography systems, existing methods still do not use continuous scene depth information, due to limitations from frame-based cameras.

Neuromorphic-based video reconstruction. Since neuromorphic sensors do not directly output image frames, all-in-focus imaging from neuromorphic cameras requires translating neuromorphic data to the domain of images. Many methods aim to reconstruct latent images that produced by neuromorphic signals for human perception, including event-based methods and spike-based methods.

Image reconstruction from events is an ill-posed problem, as events only record differential information of the scene. Reconstructing intensity frames from events can be achieved using hand-crafted features and regularization [31], [42]. More recent approaches adopt end-to-end generation methods to make use of prior knowledge. Rebecq *et al*. [34] synthesized video frames with a U-Net-like E2VID model. Weng *et al*. [53] presented a hybrid CNN-transformer network for intensity frame reconstruction. Zhu *et al*. [62] proposed a bio-inspired SNN to improve the image reconstruction quality. Gantier Cadena *et al*. [12] exploited the sparsity of event data to improve the computational efficiency of image reconstruction.

The sensors of spike cameras trigger spikes whenever accumulated photons reach a threshold. Hence, spike streams encode the scene intensity. Zhu *et al*. [61] proposed the methods "Texture From Inter-spike-intervals (TFI)" and "Texture From Playback (TFP)" to reconstruct intensity images from spike streams. In these algorithms, the temporal window sizes for reconstruction need to be carefully selected, as short windows lead to strong noise, and long windows cause motion blur. Zhao *et al*. [55] proposed using a neural network with hierarchical architecture to exploit the temporal correlation of the spike streams. Inspired by the short-term plasticity mechanism of the brain, Zheng *et al*. [58] proposed the "Texture From Motion-Dependent Short-Term Plasticity (TFMDSTP)" algorithm, which distinguishes moving and stationary regions to improve reconstruction performance.

Inspired by the ability of neuromorphic streams to capture continuous intensity changes and to reconstruct high-frame-rate videos, this paper explores how to perform focal sweeps with two types of neuromorphic cameras (either event or spike cameras) to conquer the bottlenecks of existing all-in-focus image recovery methods.

3 Proposed Method

In this section, we first introduce the neuromorphic camera formation preliminaries in Section 3.1. We then formulate the neuromorphic focal stack and our model for reconstructing refocused images in Section 3.2, and propose our general all-in-focus imaging framework in Section 3.3, and further introduce the data-driven-based method in Section 3.4 and Section 3.5. Our implementation details are illustrated in Section 3.6.

3.1 Neuromorphic camera formulation preliminaries

A neuromorphic camera operates on a fundamentally different principle compared to traditional frame-based cameras. We outline the formation models for both event camera and spike cameras as follows.

Event formulation. An event signal $e = (x, y, t, p)$ with polarity p is triggered whenever the log irradiance changes at pixel (x, y) at time t, exceeding a preset threshold c:

$$|\log(\mathbf{I}^{t}_{x,y}) - \log(\mathbf{I}^{t-\Delta t}_{x,y})| \geq c, \tag{1}$$

where $\mathbf{I}^{t}_{x,y}$ and $\mathbf{I}^{t-\Delta t}_{x,y}$ represent the pixel irradiance at coordinates (x, y) at times t and $t - \Delta t$, respectively. And the previous event of pixel (x, y) is triggered at $t - \Delta t$. Polarity $p \in \{1, -1\}$ indicates whether the intensity changes increase or decrease. Equation (1) applies to each pixel (x, y) independently, so pixel indices are omitted in the subsequent discussion. Given that the event signal detects intensity changes, most events predominantly manifest along object boundaries, as illustrated in the fourth row of Figure 2.

As events record continuous intensity changes, given two instantaneous latent images $\mathbf{I}^{t_1}$ and $\mathbf{I}^{t_2}$, assume there are N_e events occurring between t_1 and t_2, denoted as $\{e_k\}_{k=1}^{N_e}$. According to the physical model of the event camera shown in Equation (1), we can bridge $\mathbf{I}^{t_1}$ and $\mathbf{I}^{t_2}$ with corresponding events in log domain as:

$$\log \mathbf{I}^{t_2} = \log \mathbf{I}^{t_1} + \sum_{k=1}^{N_e} c_k \cdot p_k, \tag{2}$$

where c_i denotes the spatial-temporal variant threshold, related to the scene condition [15].

Spike formulation. A spike signal $s = (x, y, t)$ is triggered when the accumulated light intensity exceeds the preset thresholds ϕ:

$$\int_{t_1}^{t_2} \mathbf{A}_{x,y}(t)\mathrm{d}t \geq \phi, \tag{3}$$

in which $\mathbf{A}_{x,y}(t)$ represents the incoming light intensity at pixel location (x, y) in the time interval $[t_1, t_2]$. Upon the triggering of a spike signal, the accumulated intensity is reset. Equation (3) is also applicable to each pixel (x, y) independently; therefore pixel indices are omitted henceforth. As illustrated in the fifth row of Figure 2, there is a clear proportionality between the number of spikes and scene radiance, consistent with Equation (3). Besides, spike cameras trigger denser signals than event cameras, particularly in areas with few texture changes.

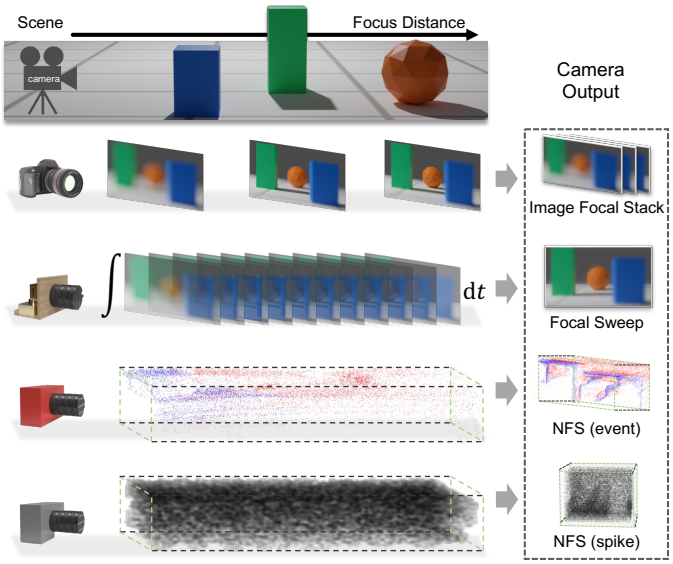


Figure 2: An illustration (from top to bottom rows) of image focal stack [59], focal sweep [20], and neuromorphic focal stack. When focus distance sweeps from near to far, each method captures information at different depths, as shown below the scene. Note that NFS (event/spike) continuously records the intensity changes and encodes texture information from all depths to which the camera focuses.

For latent image reconstruction, assume there are N_s spike occurring between t_1 and t_2, denoted as $\{s_k\}_{k=1}^{N_s}$. According the image formation model in Equation (3), the latent image $\mathbf{I}_t$ can be reconstructed with an exposure time f at any timestamp $t \in [t_1, t_2]$ as follows [61]:

$$\mathbf{I}_t = \sum_{t_k \in [I_t + f/2, I_t - f/2]} \phi_k, \tag{4}$$

where ϕ_k denotes the spatial-temporal variant threshold like the event camera.

Both events and spikes detect and record scene information using discrete signal streams, offering the potential to reconstruct images at high frame rates. Compared with frame-based cameras, neuromorphic cameras can capture more continuous scene information with a higher frame rate. This allows for neuromorphic camera applications in high-speed photography, such as motion deblurring and high-frame-rate video generation. As mentioned before, capturing a focal stack demands a high-speed camera, and it further drives our motivation to utilize neuromorphic cameras for this purpose.

3.2 Neuromorphic focal stack

As the Thin Lens Law $1/f = 1/u + 1/v$ shows (f is the focal length of the lens, u is the sensor-lens distance, and v is the object distance), we can change u or v to move the focal plane. Conventional image focal stack methods [14], [59] capture a set of images $\{\mathbf{I}^{d_i}\}^{N_d}$ with different focus distances (shown in second row of Figure 2) and then merge them with proper weights $\mathbf{W}$ to obtain an all-in-focus image, *i.e.*,

$$\mathbf{I}^{\text{AIF}} = \sum^{N_d} \mathbf{W}_i \otimes \mathbf{I}^{d_i}. \tag{5}$$

To generate an all-in-focus image, every object must have a corresponding in-focus image within the focal stack. As illustrated in the second row of Figure 2, the blue cuboid is not focused in any image of the captured focal stack, and further leads to defocus blur in the restored image. To avoid losing scene focus information in the desired depth range, the focal sweep technique [20] changes the sensor-lens distance in the exposure time, and captures an integrated defocused image, which can be seen as an all-in-focus image convolved with an integrated *Point Spread Function* (PSF), denoted as IPSF:

$$\text{IPSF}(r, u) = \int_0^T \text{PSF}(r, u, v(t))\text{d}t, \tag{6}$$

in which r represents the distance of an image point from the center of the PSF, $v(t)$ denotes the sensor-lens distance as a function of time, and T is the exposure time. Kuthirummal *et al.* [20] have proved that IPSF(r, u) is invariant to scene depth and image location to simplify the problem analysis. As the final output is a single defocused image (shown in third row of Figure 2), they also need to estimate the blur kernel and then deconvolve images, which is still an ill-posed problem.

Due to the exposure interval of conventional cameras, they cannot record high-speed changes in a scene with high fidelity, something at which neuromorphic cameras excel. Therefore, this observation inspired us to apply focal sweep to a neuromorphic camera. In fact, applying the focal sweep technique to a neuromorphic camera is quite simple. We just need to rotate the focusing ring of its lens.[1] Since the neuromorphic camera owns high temporal resolution, it outputs high-frame-rate event/spike streams that capture the pixel radiance changes as the focal plane sweeps through the scene. We call the event/spike streams during the focal sweep process as a *neuromorphic focal stack* (NFS), denoted as $\mathcal{N}$:

$$\mathcal{N} = \begin{cases} \bigcup_{u(t)} \{e_k\}_{k=1}^{N_e} \text{ (event)}, \\ \bigcup_{u(t)} \{s_k\}_{k=1}^{N_s} \text{ (spike)}, \end{cases} \tag{7}$$

where $u(t) \in (0, \infty)$ denotes the focused object distance as a function of time that transforms from nearly 0 to infinity. NFS has two important advantages over image-based focal stack (fourth and fifth row of Figure 2): 1) It continuously records the intensity changes with respect to focus distance compared with a discrete set of images [59]; 2) texture information at different depths is distinguished by the neuromorphic signal timestamps, while focal sweep method integrates the depth information [20] and only outputs a single image.

Given an image $\mathbf{I}^{d_j}$ focused at an arbitrary distance d_j, as Equation (2) and Equation (4) show, a refocused image $\mathbf{I}^{d_i}$ can be reconstructed by corresponding events/spikes. For the event-based method, we can rewrite Equation (2) to connect the refocused images as:

$$\begin{aligned} \mathbf{I}^{d_i} &= \text{C}(\mathbf{I}_{\text{L}}^{d_j} \cdot \exp(\sum_{u(t_k) \in (d_j, d_i)} c_k \cdot p_k), \mathbf{I}_{\text{ab}}^{d_j}) \\ &= \text{C}(\mathbf{I}_{\text{L}}^{d_j} \cdot \mathbf{R}^{d_j \to d_i}, \mathbf{I}_{\text{ab}}^{d_j}), \end{aligned} \tag{8}$$

1. Focal sweep setup can be found in the supplementary material.

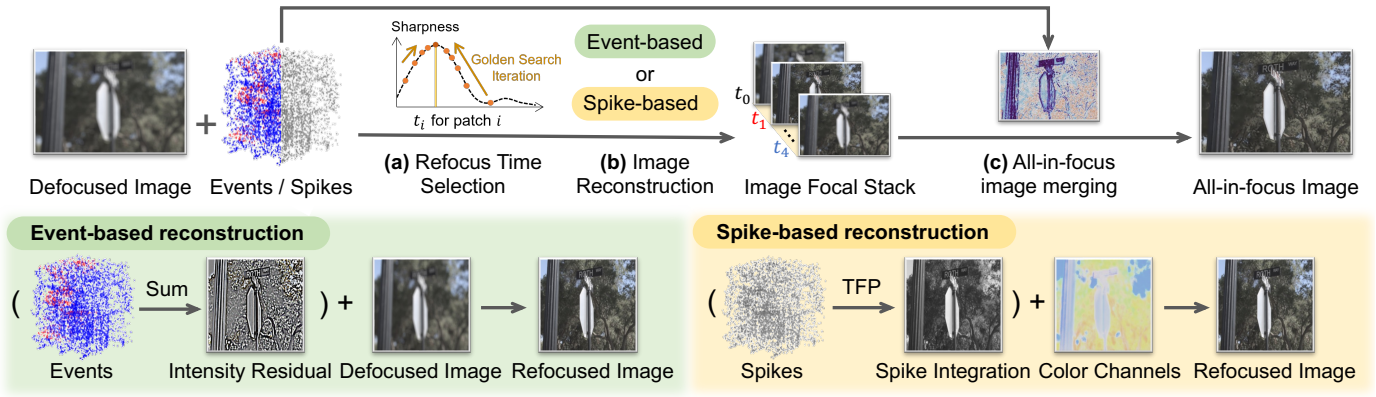


Figure 3: An illustration of our general neuromorphic-guided all-in-focus imaging pipeline. We propose the neuromorphic focal stack composed of event/spike streams, which can be used to reconstruct an image focal stack and predict the merging weights for all-in-focus image recovery. Our pipeline consists of three steps: (a) selecting the refocusing timestamps, (b) reconstructing the corresponding image focal stack, and (c) merging the stack into an all-in-focus image with weights predicted from the images and adjacent neuromorphic signals.

in which $\mathbf{I}^{d_i}$ denote the refocused images, whose focused object distances are d_j, and $\mathbf{R}^{d_j \to d_i}$ is the intensity residual computed from event summation in exponential space. $\mathbf{I}_{\mathrm{L}}^{d_j}$ and $\mathbf{I}_{\mathrm{ab}}^{d_j}$ denotes the image intensity and image color channels, respectively, while $\mathrm{C}(\cdot)$ is the color compensation operator. And for the spike-based method, we can rewrite Equation (4) to generate refocused images as:

$$\begin{aligned}\mathbf{I}^{d_i} &= \mathrm{C}\Big(\sum_{u(t_k)\in U(d_i)} \phi_k, \mathbf{I}_{\mathrm{ab}}^{d_j}\Big) \\ &= \mathrm{C}(\mathbf{G}^{d_i}, \mathbf{I}_{\mathrm{ab}}^{d_j}),\end{aligned} \tag{9}$$

where $\mathbf{G}^{d_i}$ represents spike integration within the focal distance in the neighbourhood of d_i, *i.e.*, $U(d_i) = (d_i - \delta, d_i + \delta)$.

By iteratively applying Equation (8) or Equation (9), we can obtain the image focal stack $\{\mathbf{I}^{d_i}\}^{N_d}$, consisting of N_d refocused images. As shown in Equation (5), combining the image focal stack with proper weights, an all-in-focus image can be recovered from the refocused image focal stack; thus we can obtain an all-in-focus image from the NFS and an arbitrarily focused image as inputs. Thus, the critical components of all-in-focus image restoration involve estimating the refocused timestamp and determining precise weights of Equation (5). These aspects will be elaborated in the subsequent sections.

3.3 All-in-focus imaging from NFS

The general pipeline of our method is shown in Figure 3. We first select a set of refocusing timestamps guided by reconstructed sharpness. Then we compute event-based intensity residual or spiked-based integration, which is used to reconstruct refocused images. After that, we use event or spike streams to predict merging weights and finally obtain an all-in-focus result.

Refocusing time selection. As shown in the second row of Figure 2, traditional focal stack methods [14], [59] capture a set of images with uniform time intervals. To ensure that all objects are focused in the final results, it is important that each object have a corresponding in-focus image within the focal stack, that requires a specific-designed device [59] or careful selection of refocusing distances.

According to Section 3.2, neuromorphic signals represent the scene information changes, that naturally encode the temporal gradient changes [10]. For the event-based method, assuming local events are triggered by the same edge with uniform motion, the event triggering rate is proportional to the spatial gradient. Based on such an observation, Lin *et al.* [26] designed an auto-focus algorithm for event cameras to find the maximum event triggering rate timestamp as the refocusing timestamp. However, the majority of events in the NFS are triggered by the focal sweep, instead of object motion. Thus, the event triggering rate is not suitable as a metric for refocusing timestamp selection. While for spike-based methods, there is no existing auto-focus algorithm.

To obtain an accurate refocusing timestamp in both event-based and spike-based methods, we do not search it in the neuromorphic signal domain. Inspired by the image-based auto-focus method [49], we use reconstructed image sharpness as a focus metric. We fuse the NFS with a given RGB image to reconstruct refocused images by Equation (8) or Equation (9), and then utilize the variance of reconstructed image intensity value $\mathbb{D}(\mathbf{I})$ to evaluate the image sharpness. We assume that the time t_r with the maximum variance value is the refocusing timestamp we want to find. We adopt the golden-section search method in [18] to NFS for searching the time t_r with maximal image sharpness, as summarized in Algorithm 1.

The depths of objects in a scene are different, leading to the different refocusing time. Therefore, we split the image into $N \times N$ spatially non-overlapping patches $\{\mathbf{I}_p^{d_j}\}^{N\times N}$, with corresponding NFS patches $\{\mathcal{E}_p\}^{N\times N}$. We apply the aforementioned Algorithm 1 to each of patches to find their refocusing times, resulting in a set of $N \times N$ refocusing timestamp, *i.e.*,

$$\{t_r\}^{N\times N} = \bigcup_p \mathtt{TS}(\mathcal{N}_p, \mathbf{I}_p^{d_j}), \tag{10}$$

where $\mathtt{TS}$ denotes refocusing time selection with NFS using the golden-section search method [18].

Algorithm 1 Refocusing time selection with NFS

Data: threshold μ, golden ratio $\varphi = 1.618$
Input: NFS $\mathcal{N}$ and an RGB image $\mathbf{I}^d$
Result: Refocusing timestamp t_r
$L \leftarrow 0, R \leftarrow N$
while $R - L > \mu$ **do**
  $t_1, t_2 \leftarrow R - (R-L)/\varphi, L + (R-L)/\varphi$
  Reconstruct $\mathbf{I}^{d_1}, \mathbf{I}^{d_2}$ with Equation (8)/ Equation (9)
  if $\mathbb{D}(\mathbf{I}^{d_1}) > \mathbb{D}(\mathbf{I}^{d_2})$ **then** $R \leftarrow t_2$
  else $L \leftarrow t_1$
  end if
  $t_r \leftarrow (L + R)/2$
end while

Refocused image reconstruction. After the refocused timestamp is selected, with the guidance of neuromorphic signals, we can reconstruct a set of refocused images. As depicted in Equation (8) and Equation (9), the crux of the method lies in estimating the threshold for neuromorphic cameras. When it comes to event-based refocused image reconstruction, estimating the threshold for an event camera poses a challenge. Drawing inspiration from Pan *et al*. [32], we directly reconstruct the refocused images and employ the subsequent energy function to appraise the reconstructed refocused images:

$$\mathscr{E}(\mathbf{I}) = -\alpha \cdot \texttt{Sobel}(\mathbf{I}) + \beta \cdot \texttt{TV}(\mathbf{I}), \tag{11}$$

in which $\mathscr{E}$ is the energy function, the Sobel loss $\texttt{Sobel}(\cdot)$ is the average value of image $\mathbf{I}$ after using a Sobel filter [46], and the TV loss [40] $\texttt{TV}(\cdot)$ is the total variation of the image. The Sobel loss promotes pronounced edges, while the TV loss advocates for smoothness and noise mitigation. Minimizing this function allows us to derive the image focal stack.

For spike-based reconstruction, we employ the "Texture From Playback" (TFP) approach as presented in [61]. This approach reconstructs the image by aggregating spike counts over a brief time span. Due to the corruption from the leaking current, the thresholds of spike cameras are also not constant. We combine the RGB defocus image with corresponding spikes to estimate this threshold for each signal. By using the inferred threshold, we can compute spike integration to approximate the gray-scale intensity.

Note that both Equation (8) and Equation (9) are applied to the intensity channel. Thus, we convert the RGB image into the Lab color space and first process intensity channel (L). By combining the intensity channel, reconstructed from event/spike streams, with the respective color channels (ab), we can produce the refocused images with colors. Thanks to the continuous information encoded by neuromorphic signal streams, we can refocus images to any arbitrary time.

All-in-focus image merging. The quality of a merged all-in-focus image, derived from an image focal stack, is fundamentally contingent upon the precise estimation of merging weights. An approach to merging images within a focal stack, by harnessing image spatial gradient as a guide, is discussed by Horn [14]. As described in Equation (1) and Equation (3), neuromorphic signal streams inherently capture essential information along edges. These inherent traits serve as reliable cues for the prediction of merging weights.

(a) Defocused Image (b) Ground Truth (c) Events (d) Spikes

Figure 4: An example results of all-in-focus imaging with the non-learning method. (a) Defocused Image. (b) Ground Truth. (c) and (d) All-in-focus images recovered using events and spikes, respectively. Due to the noise effects, this direct fusion approach often leads to undesirable ringing artifacts (highlighted in green boxes), and the recovered images still remain blurry in text-rich regions (highlighted in orange boxes). Close-up views are provided at the bottom left of each image.

When considering event-based weight computation, it is discernible that as the focal point traverses through a point within a scene, the blur kernel subjected to it oscillates from broad to narrow and broadens again—indicating a symmetric operation. Consequently, the corresponding pixel value undergoes a symmetric alteration. Thus, every event precipitated prior to the in-focus moment aligns with a subsequent event of an opposing polarity post the in-focus moment. Grounded in this observation, Bao et al. [4] assessed the distribution of all positive and negative events chronologically, pinpointing the in-focus time t that yields symmetry between positive and negative distributions, with t serving as the axis of symmetry. For computing the in-focus times at a per-pixel level, we amplify this method. This amplification is essential since the volume of events triggered on a lone pixel is insufficient to determine distributions with robustness. To gauge the likelihood of time t_i as the in-focus time for a pixel, we introduce the metric $f(t_i)$ as delineated:

$$\begin{aligned} f^1(t_i) &= \min(\texttt{CNT}(t_i, 0, \leq), \texttt{CNT}(t_i, 1, \geq)), \\ f^2(t_i) &= \min(\texttt{CNT}(t_i, 1, \leq), \texttt{CNT}(t_i, 0, \geq)), \\ f(t_i) &= \max(f^1(t_i), f^2(t_i)), \end{aligned} \tag{12}$$

where $\texttt{CNT}(t, p, \leq \ / \ \geq)$ signifies the count of events with triggering times lesser/greater than t, contingent upon the polarity p. To culminate, we deploy a Gaussian filter to seamlessly smooth $f(t)$ across the image.

For the spike-based weight calculation, spikes encode relative intensity values via their trigger frequency. Consequently, we can designate the triggering frequency within

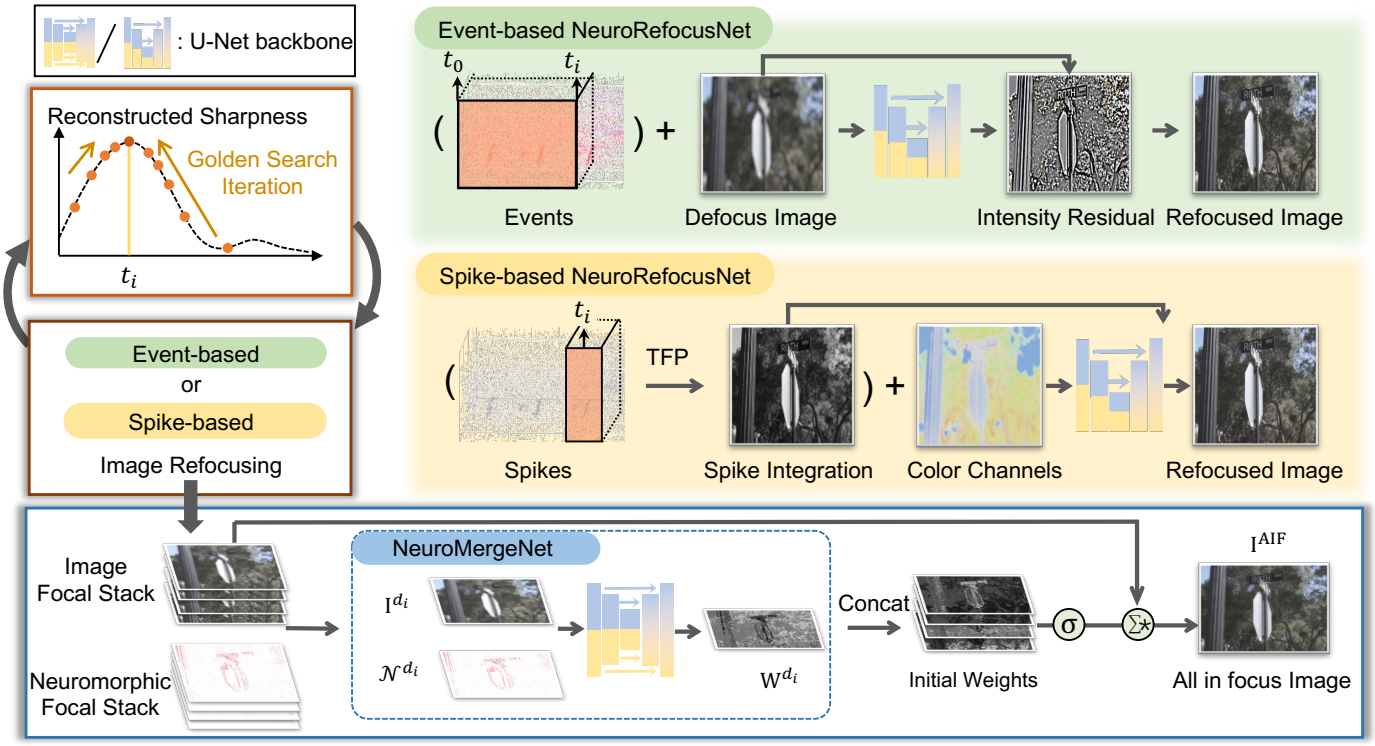


Figure 5: The pipeline of our data-driven method. We first iteratively find a refocusing timestamp according to the reconstructed sharpness for each image patch. NeuroRefocusNet fuses two modalities of data (RGB image $\mathbf{I}^{d_j}$ and NFS $\mathcal{N}^{d_j \to d_i}$) to reconstruct a refocused image $\mathbf{I}^{d_i}$. For the event-based method, NeuroRefocusNet predicts intensity residual, while for the spike-based method, it merges spike integration and color channels. By applying NeuroRefocusNet on each timestamp, N refocused images are generated, forming an image focal stack. Then, NeuroMergeNet predicts initial weights guided by the neuromorphic signals in the related time interval of each refocusing timestamp and concatenates initial weights. Merging weights are transformed from initial weights by the softmax function. Finally, the reconstructed image focal stack is merged the with corresponding predicted weights to obtain an all-in-focus result $\mathbf{I}^{\mathrm{AIF}}$.

a stipulated time window of image exposure time as the metric, represented as:

$$f(t_i) = \mathbb{D}_{t_k \in U(t_i)}\{s_k\}. \tag{13}$$

To determine the merging weights, the derived metric value is normalized and subsequently deployed for the recuperation of an all-in-focus image.

Utilizing this all-in-focus image merging pipeline allows us to reconstruct all-in-focus images. An illustrative example can be seen in Figure 4. As the results show, the fusion method introduces some unwanted artifacts, leaving the recovered images with lingering blurriness, especially in areas rich in text. As detailed in Equation (2) and Equation (4), the threshold of a neuromorphic camera is not fixed [15], [25], [61]. Furthermore, neuromorphic cameras are prone to current leakage [15], [25], [61], causing noisy signals that fluctuate based on lighting conditions. The process of determining merging weights from these neuromorphic signals can be compromised by this noise, subsequently affecting the overall quality of the reconstructed images. To address these challenges, we further introduce data-driven approaches in subsequent stages.

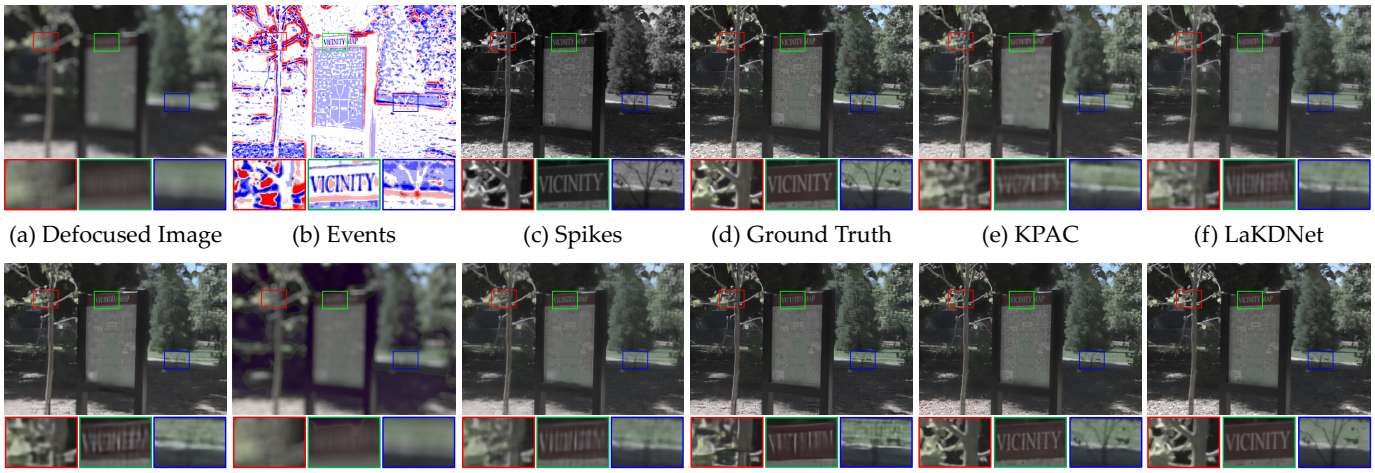
3.4 NeuroRefocusNet

Given a refocusing timestamp, attempting a direct reconstruction of a refocused image using Equation (8) and Equation (9) with a static threshold results in pronounced artifacts. To address the challenges posed by spatial-temporal variant thresholds, we propose a U-Net architecture network, which we call NeuroRefocusNet. This is designed to predict the residual between the input defocused image and the resulting refocused image in a data-driven approach. For event-based reconstruction, this model captures $\mathbf{R}$ as illustrated in Equation (8), and it also accounts for noise effects in spike-based reconstructions. With a collection of refocusing timestamps determined from Algorithm 1, we derive an image focal stack from an RGB image and the NFS, expressed as:

$$\{\mathbf{I}^{d_i}\}^{N\times N} = \mathtt{f}_{\mathrm{r}}(\mathbf{I}^{d_j}, \{\mathcal{N}^{d_j \to d_i}\}^{N\times N}, \{t_r\}^{N\times N}), \tag{14}$$

where $\mathbf{I}^{d_j}$ is a given RGB image focusing at an arbitrary distance d_j, $\mathcal{N}^{d_j \to d_i}$ denotes corresponding neuromorphic signals triggering between $\mathbf{I}^{d_j}$ and $\mathbf{I}^{d_i}$, and $\mathtt{f}_{\mathrm{r}}$ is an implicit function modeled by NeuroRefocusNet. As input images represent scene conditions to some extent (the defocused regions are blurry), the network can predict residual with spatial-temporal variant thresholds guided by input images.

The efficacy of the multi-scale architecture in multi-modal data fusion has been well-documented in [48]. Consequently, we employ the U-Net model for both event-based and spike-based reconstruction. As depicted in Figure 5, for event-based reconstruction, we merge the image and NFS



(a) Defocused Image (b) Events (c) Spikes (d) Ground Truth (e) KPAC (f) LaKDNet

(g) IFAN (h) APL (i) DRBNet (j) Restormer (k) Ours (ev) (l) Ours (sp)

Figure 6: Visual quality comparison with image-based defocus deblurring methods on synthetic data. (a) Defocused Image. (b) Events. (c) Spikes. (d) Ground Truth. (e)∼(l) All-in-focus results of KPAC [47], LaKDNet [37], IFAN [22], APL [57], DRBNet [38], Restormer [54], ours (ev) and ours (sp). More results are in the supplementary material.

features at multiple scales using a U-Net backbone[2]. This is then formulated through residual learning with global connections. By superimposing this residual onto the input RGB image, we can successfully restore the refocused images. Meanwhile, for spike-based reconstruction, we use the initial refocused images derived from the TFP methods [61] as inputs. These are then refined for quality using a U-Net backbone, effectively eliminating noise and color bleeding.

3.5 NeuroMergeNet

As elaborated in Section 3.3, extracting weights directly from neuromorphic streams is impeded by signal noise. Given that the sharpness of each pixel can be directly discerned from the image focal stack, superior-quality merging weights are derived from combining the image focal stack and the NFS.

As depicted in Figure 5, we introduce another U-Net architecture [36] dubbed NeuroMergeNet, to predict initial weights for each image, represented as:

$$\{\mathbf{W}^{d_i}\}_{\text{init}} = \mathtt{f}_{\text{m}}(\{\mathbf{I}^{d_i}\}^{N\times N}, \mathcal{N}^{d_i}). \tag{15}$$

In this formula, $\{\mathbf{W}^{d_i}\}_{\text{init}}$ symbolizes the weight matrix set, each with dimensions $H \times W$, and $\mathtt{f}_{\text{m}}$ stands for an implicit function embodied by NeuroMergeNet. Within the initial weights $\{\mathbf{W}^{d_i}\}_{\text{init}}$, a higher value indicates sharper pixel intensity. For determining the final merging weights, the initial weights undergo a transformation via the softmax function, represented as:

$$\mathbf{W} = \texttt{Softmax}(\alpha \cdot \{\mathbf{W}^{d_i}\}_{\text{init}}), \tag{16}$$

where α is a hyper-parameter, its elevated value diminishing the influence of smaller noisy values.

Given that NeuroMergeNet produces initial weights for each image individually, the set number of the image focal stack can be unfixed. This flexibility means our NeuroMergeNet can accommodate an image focal stack of any size. After calculating the initial weights for each image, these sets of weights are combined to derive the final weights. Furthermore, to circumvent over-fitting on synthetic data, we pay more attention to predicting merging weights, rather than directly restoring an all-in-focus image. Adhering to the merging procedure outlined in [59], we consistently apply the same weight map to all three RGB channels and merge them independently using refocused images.

3.6 Implementation details

Dataset. Since there is no large-scale image focal stack dataset with event information, we render a synthetic image focal stack with Blender [5] and simulate corresponding event/spike streams with the latest simulator. We choose DVS-Voltmeter [25] and Spikesim [56] for event and spike simulation, respectively. Our dataset is composed of 200 random scenes. From each scene, we render an image focal stack with a shallow DoF camera setup (aperture f/1.2, focal length 100mm), which sweeps its focus distances through the scene, and an all-in-focus image as ground truth. We further scale and scatter geometric objects to increase the diversity. To better match the data distribution to real-world images, we wrap the surfaces of the objects with images sampled from the MS-COCO dataset [27] as their textures. After rendering the image focal stack with 480 frames, we input them into DVS-Voltmeter [25]/Spikesim [56] to generate event/spike streams[3].

Training details. Both NeuroRefocusNet and NeuroMergeNet are trained with the same loss function as:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{perc}}(\mathbf{I}^{\text{o}}, \mathbf{I}^{\text{gt}}) + \beta \cdot \mathcal{L}_2(\mathbf{I}^{\text{o}}, \mathbf{I}^{\text{gt}}), \tag{17}$$

where $\alpha = 0.5$, $\beta = 200$, $\mathcal{L}_2$ denotes the MSE loss, and $\mathcal{L}_{\text{perc}}$ denotes a perceptual loss calculated from a VGG-19 network [45] pre-trained on ImageNet [41]. The output image $\mathbf{I}^{\text{o}}$ is

2. Detailed network configurations can be found in the supplementary.

3. More details about the data generation pipeline can be found in the supplementary material.

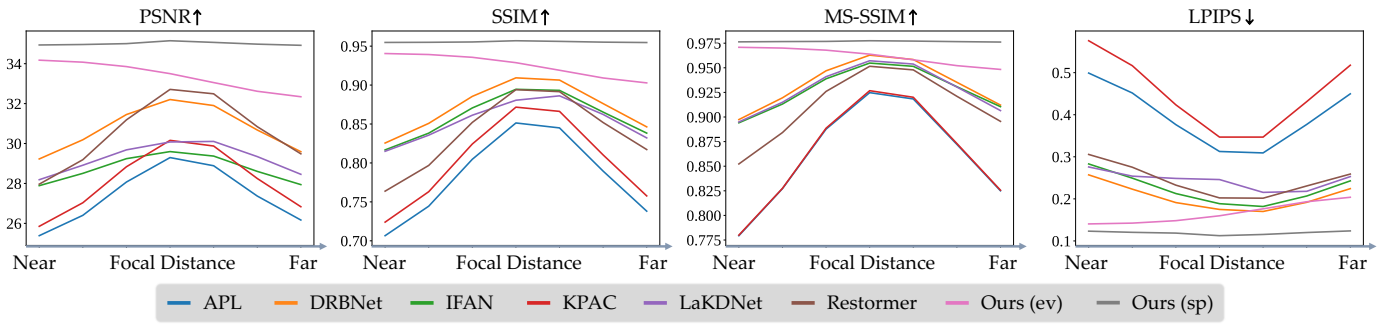


Figure 7: Quantitative comparison across different focal distances on the LiFF dataset [7]. We input defocused images spanning a range of focal distances (from near to far) and plot the averaged results for each focal distance across all scenes. ↑ (↓) indicates that higher (lower) is better. Our method consistently outperforms other methods at all focal distances, demonstrating stable performance.

Table 1: Quantitative comparisons on the synthetic LiFF dataset [7]. ↑ (↓) indicates the higher (lower), the better throughout this paper.

	PSNR↑	SSIM↑	MS-SSIM↑	LPIPS↓
KPAC [47]	26.00	0.7643	0.8402	0.3778
IFAN [22]	26.97	0.7891	0.8644	0.3435
APL [57]	24.33	0.6753	0.7158	0.5471
DRBNet [38]	27.75	0.7882	0.8583	0.3243
Restormer [54]	26.43	0.7210	0.7979	0.3472
LaKDNet [37]	27.11	0.7830	0.8629	0.3324
Ours (ev)	33.25	0.9323	0.9611	0.1510
Ours (sp)	34.85	0.9541	0.9768	0.1278

the predicted refocused image or all-in-focus image. The corresponding ground truth is denoted as $\mathbf{I}^{gt}$.

We implement our method with PyTorch on a single NVIDIA GeForce RTX 3090 Ti GPU. We train both NeuroRefocusNet and NeuroMergeNet for 100 epochs, starting with the learning rate 5×10^{-4}, and after the first 50 epochs, we decrease the learning rate by $1/10$ for every 20 epochs. The ADAM optimizer [19] is used in the training phase. For event-based NeuroRefocusNet training, we randomly select two frames from our synthetic dataset, one as input and the other as ground truth. With corresponding events triggered between them, NeuroRefocusNet can reconstruct refocused images. For spike-based NeuroRefocusNet training, we need to input an image and part of spikes to reconstruct refocused images. For the input of NeuroMergeNet, we input all generated refocused images and the corresponding event/spike streams for guidance.

4 Experimental Results

In this section, we qualitatively and quantitatively compare our method with state-of-the-art image-based defocus deblurring methods on a public synthetic dataset (Section 4.1) and our real-captured data (Section 4.2). In addition, we illustrate the advantages and disadvantages of employing event cameras in comparison to spike cameras for all-in-focus image recovery (Section 4.3). In Section 4.4, ablation studies are conducted to validate the effectiveness of each module of the proposed method.

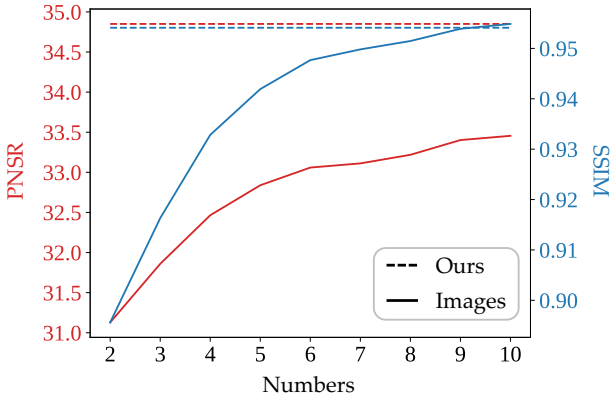


Figure 8: A quantitative comparison using varying numbers of images in the focal stack method. The number of images in the focal stack ranges from 2 to 10 frames. The red curves represent the results of PSNR, and blue curves for SSIM, respectively. Capturing more than 10 frames marginally surpasses our method in terms of the SSIM metric.

4.1 Quantitative comparison using synthetic data

As the majority of existing all-in-focus image datasets do not contain image focal stacks, we generate image focal stacks from a light field dataset, the Stanford Multiview Light Field (LiFF) Dataset [7], which was captured with hand-held Lytro Illum cameras. We synthesize the image focal stacks from their light field images and corresponding paired events/spikes are generated by DVS-Voltmeter [25]/Spikesim [56]. The first image of each focal stack is selected as the input defocused image. Among all the synthetic triplet clips, consisting of defocused images, all-in-focus images, and corresponding neuromorphic cameras, we select 50 sets that are consistent with the LFDOF dataset [39] as our testing dataset for a fair comparison with other methods.

We compare our method with six recent image-based defocus deblurring methods: DRBNet [38], IFAN [22], KPAC [47], APL [57], Restormer [54], and LaKDNet [37]. We also denote our method incorporating event cameras as ours (ev) while ours (sp) for spike cameras. The quantitative comparisons are shown in Table 1 and qualitative comparisons are shown in Figure 6. Our method outperforms other state-of-the-art methods with more than 15% improvement on three metrics (PSNR, SSIM, LPIPS) by event cameras,

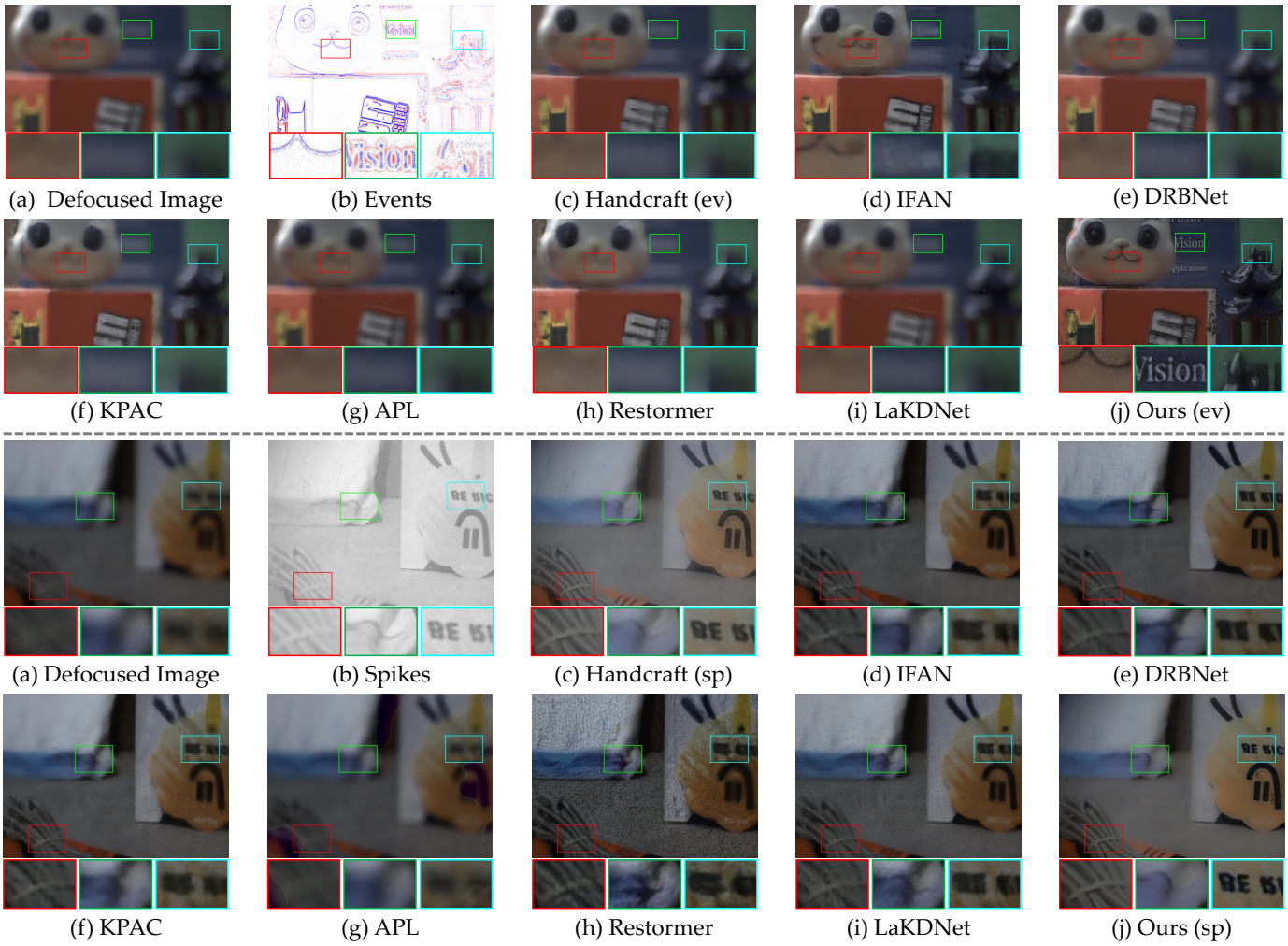


Figure 9: Visual quality comparison with image-based defocus deblurring methods on real data. (a) Defocused image. (b) Event / Spike. (c)∼(j) All-in-focus results of Handcraft(ev) / Handcraft (sp), IFAN [22], DRBNet [38], KPAC [47], APL [57], Restormer [54], LaKDNet [37], and ours (ev) / ours (sp). More results are in the supplementary material.

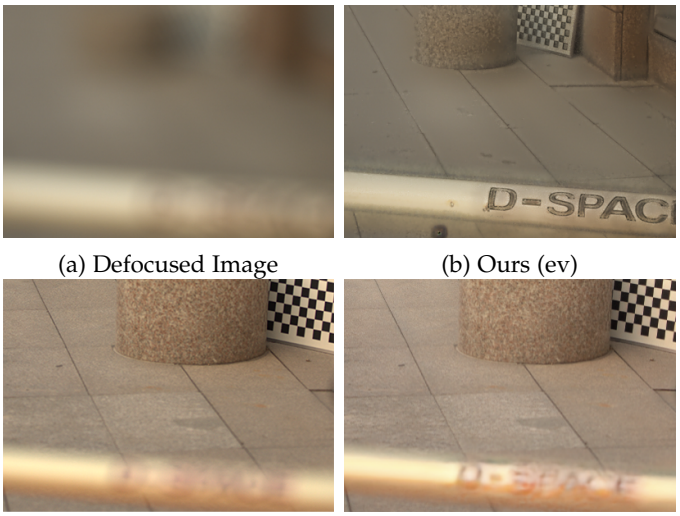
while more than 20% improvements for spike cameras, restores more high-frequency details encoded inside the neuromorphic streams, and recovers an all-in-focus image with higher quality and fewer artifacts.

Note that this comparison might seem somewhat biased, as image-based defocus methods rely solely on a single image input. Our purpose is to show that a significant performance boost can be achieved when continuous information from neuromorphic streams is involved. Additionally, we conduct an experiment to assess the influence of defocus blur level at varying focal distances, with the results presented in Figure 7. These results demonstrate that image-based methods are notably sensitive to changes in focal distance and struggle to effectively address severe defocus blur. In contrast, by leveraging the high-temporal resolution information inherent in the NFS, our approach is capable of reconstructing an all-in-focus image from any defocused input while maintaining stable performance.

With image-based focal stack methods. Our method requires only a single exposure time to capture all-in-focus images. We conduct experiments to explore the comparative efficiency of using different numbers of images in an image focal stack. We merge all-in-focus images with varying numbers of input images, and the results are shown in Figure 8. As illustrated in Figure 8, capturing more than 10 frames slightly surpasses our method in terms of the SSIM metric.

4.2 Qualitative comparison using real data

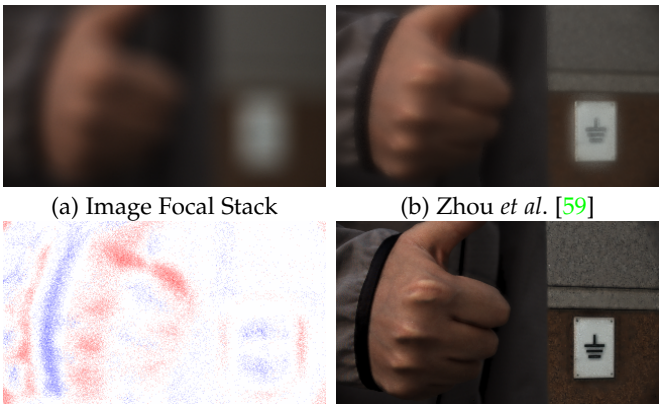
To verify the effectiveness of our method in real-world scenarios, we capture real data by building a hybrid camera system, which consists of a machine vision camera (HIKVISION MV-CA050-12UC) and a neuromorphic camera with a beam splitter. Within this setup, we selected the PROPHESEE GEN4.0 as our event camera and the Spike Camera-001T-Gen2 as our spike camera. We synchronously capture an NFS and an RGB image focused at an arbitrary distance in both indoor and outdoor scenarios. For calibration, we use a checkerboard to deal with homography and radial distortion between two views. Visual quality comparisons of all-in-focus results are shown in Figure 9. Our method can recover all-in-focus images with the correct texture in defocused regions. In comparison, other image-based methods cannot recover the sharp details well and even introduce undesired ringing artifacts.



(a) Defocused Image (b) Ours (ev)

(c) Dual-pixel Image (d) DPDNet [1]

Figure 10: Visual quality comparison with a dual-pixel method. (a) Defocused image captured by machine vision camera. (b) All-in-focus images recovered by our method. (c) Dual-pixel image captured by Canon 5D Mark IV. (d) All-in-focus image recovered by DPDNet [1] using (c).



(a) Image Focal Stack (b) Zhou *et al*. [59]

(c) Events (d) Ours (ev)

Figure 11: Visual quality comparison with an image-based focal stack method. (a) Image focal stack. (b) All-in-focus image restored by Zhou *et al*. [59]. (c) The visualization of Events. (d) All-in-focus image restored by ours (ev). Please check the supplementary video for animated results.

With dual-pixel methods. Dual-pixel images are validated as effective inputs to recover an all-in-focus image [1], [2]. These methods also take additional input like our method. To compare the performance between NFS and dual-pixel imaging, we capture a real scenario with our hybrid camera system and a Canon 5D Mark IV DLSR camera. Since the quantitative results have shown spike cameras have better reconstruction ability, we choose an event camera for comparison in this scenario without loss of generality (the same as compared with image-based focal stack methods). We compare with dual-pixel-based defocus deblurring method DPDNet [1], and the results are shown in Figure 10. We can see our method outperforms DPDNet [1]. Thanks to the high temporal resolution information in neuromorphic streams, our method recovers clearer texture information. Since the DLSR camera cannot be directly mounted on our beam splitter and the lenses are also different, we cannot obtain the NFS and dual-pixel image with perfect spatial alignment. Thus the field of view and DOF in this example are somewhat inconsistent, but the levels of details recovered by these two methods are clearly different.

With image-based focal stack methods. Traditional image-based focal stack methods [3] require capturing multiple images at different focus distances, which are sensitive to camera shake. Although Zhou *et al*. [59] proposed a space-time refocusing method to stabilize the input images by selecting corresponding pixels in the focal stack, they still require that the velocity of focal sweep is constant, which limits the applicability of their method. Since we rotate the lens to capture the image/neuromorphic focal stack, leading to unavoidable camera shake, we show that our method is robust to such slight motion and produces a sharper all-in-focus image, shown in Figure 11, while the result of Zhou *et al*. [59] shows ringing artifacts. Moreover, all real data are captured using the same setup, emphasizing the robustness of this method to minor camera shakes that occur during the rotation of the focus ring (*e.g*., Figure 9).

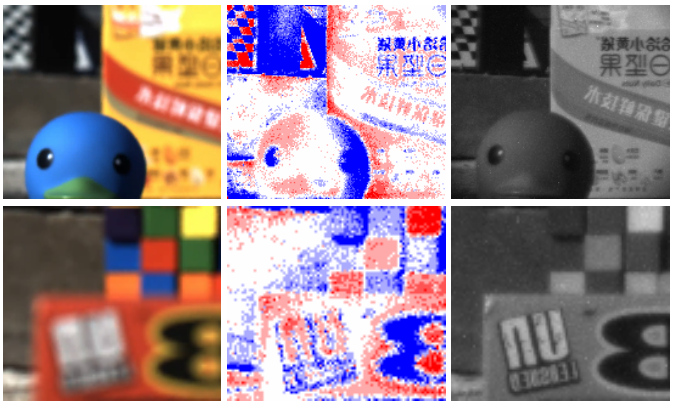
4.3 Comparison between event and spike cameras

While both event and spike cameras offer valuable assistance in all-in-focus image restoration, spike signals provide richer texture information than event cameras, as illustrated in Figure 2. The richer texture information in spike signals yields better quantitative results than events, as demonstrated in Table 1. For visual comparison in Figure 6, the results from spikes reveal more accurately reconstructed details. Furthermore, spike signals enable the direct reconstruction of grayscale all-in-focus images, a capability not present in event signals.

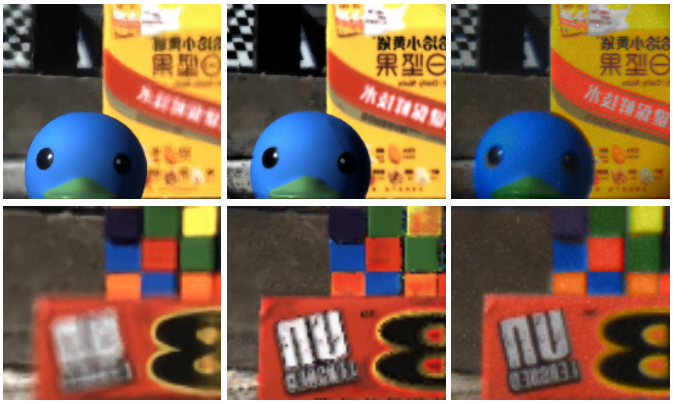
We capture these two types of data under the same scenario[4], and results are shown in Figure 12. The results demonstrate that the spike camera preserves better fidelity in textureless regions, so that the recovered images generally look clearer. The event camera keeps sharper edges, but contain artifacts in smooth regions. Another observation is that the results from the event camera seem to have better color appearance. This is because the spike-based solution needs to merge with UV channels from defocused RGB images with color distortion, while event-based solution avoids this issue by only merging intensity changes from the defocused RGB images.

Moreover, the data size difference between the two types of neuromorphic data cannot be ignored. In our test dataset, event signals average around 21 Megabytes per scenario, whereas spike signals are around 240 Megabytes. This difference requires a higher data transfer bandwidth for spike signals. Additionally, event cameras own a higher temporal resolution (10^{-6}s versus 10^{-4}s) compared to spike cameras. Such attributes make event cameras more effective in dealing with scenes with faster motion but with few textureless regions.

4. Detailed setup can be found in the supplementary material.



(a) Defocused Image (b) Events (c) Spikes



(d) LaKDNet [37] (e) Ours (ev) (f) Ours (sp)

Figure 12: Visual comparison of all-in-focus results from event and spike cameras. (a) Defocused Image. (b) Events. (c) Spikes. (d)∼(f) All-in-focus results of LaKDNet [39], ours (ev), and ours (sp).

In summary, each camera type has its strengths and weaknesses for all-in-focus image recovery. In scenarios with data bandwidth constraints, event cameras outperform spike cameras. However, spike cameras provide more reliable reconstruction quality. The current design strategy balances these trade-offs for different camera choices. Exploring methods that effectively utilize both types of neuromorphic information remains an area for future work.

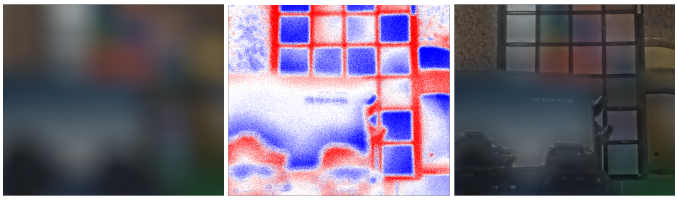
4.4 Ablation studies

To verify the effectiveness of our proposed method in cross-modality data fusion, we present the quantitative results of hand-crafted methods in Table 2 and qualitative comparison on Figure 9. Here, "Handcraft (ev)" denotes fusion with event data, and "Handcraft (sp)" refers to fusion with spike data. As the results indicate, the hand-crafted method is unable to fully exploit the high-temporal resolution information contained within the neuromorphic signals due to the presence of noise. Notably, event data exhibit a larger domain gap with images compared to spike data. Consequently, adopting a data-driven approach to bridge this domain gap between the two modalities is crucial. Through this, we can generate high-quality all-in-focus images.

To validate the efficacy of each component of our methodology, we undertake a series of ablation studies, with results presented in Table 2. Given that our network architecture is consistent for both event and spike cameras, ablation studies are conducted on one type of neuromorphic camera. For this comparative analysis, we select the event camera. We show the effectiveness of NeuroRefoucsNet by replacing it with ET-Net [53], an event-based image reconstruction method (denoted as "ET+MNet"). We further verify the contribution of NeruoMergeNet compared with all-in-focus imaging from gradient domain fusion [59] (denoted as "RNet+GDF"). Finally, we demonstrate the necessity of refocusing time selection by substituting it with uniform time selection (denoted as "Uniform"). As the results show, our complete model achieves the best performance.

Table 2: Quantitative results of ablation study.

	PSNR↑	SSIM↑	MS-SSIM↑	LPIPS↓
Handcraft (ev)	23.55	0.7589	0.8022	0.3738
Handcraft (sp)	26.40	0.8037	0.8634	0.3415
ET+MNet	12.62	0.3474	0.2169	0.7179
RNet+GDF	32.66	0.9272	0.9556	0.1698
Uniform	32.84	0.9224	0.9564	0.1605
Ours (ev)	33.25	0.9323	0.9611	0.1510



(a) Defocused Image (b) Events (c) Ours (ev)

Figure 13: Failure case: Recovering an all-in-focus image from a significantly defocused color checkerboard.

5 Conclusion

In this paper, we propose a novel neuromorphic focal stack to record intensity changes with respect to focus distance, which compacts with two types of neuromorphic cameras, compared to our previous work [29]. With NFS, we introduce a reliable refocusing timestamp selection algorithm, and further design NeuroRefocusNet and NeuroMergeNet to recover an all-in-focus image. Thanks to successfully exploring the continuous focusing related information from NFS, our method exhibits superior performance over state-of-the-art methods.

Limitations. In our present hybrid camera system, our event/spike cameras capture only gray-scale changes in intensity, thus failing to reflect differences across the RGB channels. As illustrated in Figure 13, although our method can restore sharp edges of color checkerboards, but cannot accurately recover the color space. This discrepancy is particularly pronounced when the input defocused image is significantly blurring. Consequently, our network faces challenges in compensating for a precise color map. In our current approach, we still obtain an NFS by manually rotating the lens, which is not enough to capture scenes involving

rapid object movements. Therefore, our technique is not well-suited for dynamic scenes. One potential avenue for improvement involves integrating a rapid focusing mechanism, such as a liquid lens [28], to better accommodate scenes with motion, which is left as our future work.

ACKNOWLEDGMENT

This work was supported by National Science and Technology Major Project (Grant No. 2021ZD0109803), Beijing Natural Science Foundation (Grant No. L233024), and National Natural Science Foundation of China (Grant No. 62136001, 62088102).

REFERENCES

[1] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *Proc. of European Conference on Computer Vision*, 2020. 3, 11
[2] Abdullah Abuolaim, Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Learning to reduce defocus blur by realistically modeling dual-pixel data. In *Proc. of International Conference on Computer Vision*, 2021. 3, 11
[3] Aseem Agarwala, Mira Dontcheva, Maneesh Agrawala, Steven Drucker, Alex Colburn, Brian Curless, David Salesin, and Michael Cohen. Interactive digital photomontage. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, pages 294–302, 2004. 11
[4] Yuhan Bao, Lei Sun, Yuqin Ma, Diyang Gu, and Kaiwei Wang. Improving fast auto-focus with event polarity. *Opt. Express*, 31(15):24025–24044, Jul 2023. 6
[5] Blender Foundation. The Blender project - free and open 3D creation software. Accessed: 2022-11-04. 8
[6] Shoushun Chen and Menghan Guo. Live demonstration: CeleX-V: A 1m pixel multi-mode event-based sensor. In *Proc. of Computer Vision and Pattern Recognition Workshops*, 2019. 2
[7] Donald G. Dansereau, Bernd Girod, and Gordon Wetzstein. LiFF: Light field features in scale and depth. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 9
[8] Edward R Dowski and W Thomas Cathey. Extended depth of field through wave-front coding. *Applied Optics*, 34(11):1859–1866, 1995. 3
[9] 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
[10] Peiqi Duan, Zihao W Wang, Boxin Shi, Oliver Cossairt, Tiejun Huang, and Aggelos K Katsaggelos. Guided event filtering: Synergy between intensity images and neuromorphic events for high performance imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8261–8275, 2021. 5
[11] Laurent D'Andrès, Jordi Salvador, Axel Kochale, and Sabine Süsstrunk. Non-parametric blur map regression for depth of field extension. *IEEE Transactions on Image Processing*, 25(4):1660–1673, 2016. 1, 2
[12] Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. Sparse-e2vid: A sparse convolutional model for event-based video reconstruction trained with real event noise. In *Proc. of Computer Vision and Pattern Recognition Workshops*, 2023. 3
[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proc. of International Conference on Computer Vision*, 2017. 1
[14] Berthold Klaus Paul Horn. Focusing. Technical report, MIT, 1968. 2, 3, 4, 5, 6
[15] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic DVS events. In *Proc. of Computer Vision and Pattern Recognition Workshops*, 2021. 3, 7
[16] 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. 2
[17] Ali Karaali and Claudio Rosito Jung. Edge-based defocus blur estimation with adaptive scale selection. *IEEE Transactions on Image Processing*, 27(3):1126–1137, 2017. 1, 2
[18] Jack Kiefer. Sequential minimax search for a maximum. In *Proc. of the American Mathematical Society*, 1953. 5
[19] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 9
[20] Sujit Kuthirummal, Hajime Nagahara, Changyin Zhou, and Shree K Nayar. Flexible depth of field photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):58–71, 2010. 3, 4
[21] Junyong Lee, Sungkil Lee, Sunghyun Cho, and Seungyong Lee. Deep defocus map estimation using domain adaptation. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 2
[22] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 1, 2, 8, 9, 10
[23] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 26(3):70, 2007. 3
[24] Anat Levin, Samuel W Hasinoff, Paul Green, Frédo Durand, and William T Freeman. 4D frequency analysis of computational cameras for depth of field extension. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 28(3):1–14, 2009. 3
[25] Songnan Lin, Ye Ma, Zhenhua Guo, and Bihan Wen. DVS-Voltmeter: Stochastic process-based event simulator for dynamic vision sensors. In *Proc. of European Conference on Computer Vision*, 2022. 7, 8, 9
[26] Shijie Lin, Yinqiang Zhang, Lei Yu, Bin Zhou, Xiaowei Luo, and Jia Pan. Autofocus for event cameras. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 5
[27] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. of European Conference on Computer Vision*, 2014. 8
[28] Carlos A López and Amir H Hirsa. Fast focusing using a pinned-contact oscillating liquid lens. *Nature Photonics*, 2(10):610–613, 2008. 13
[29] Hanyue Lou, Minggui Teng, Yixin Yang, and Boxin Shi. All-in-focus imaging from event focal stack. In *Proc. of Computer Vision and Pattern Recognition*, 2023. 2, 12
[30] Marcella Matrecano, Melania Paturzo, and Pietro Ferraro. Extended focus imaging in digital holographic microscopy: a review. *Optical Engineering*, 53(11):112317, 2014. 1
[31] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 126(12):1381–1393, 2018. 3
[32] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 6
[33] Jinsun Park, Yu-Wing Tai, Donghyeon Cho, and In So Kweon. A unified approach of multi-scale deep and hand-crafted features for defocus estimation. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 2
[34] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):1964–1980, 2019. 2, 3
[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Adv. in Neural Information Processing Systems*, 2015. 1
[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. of International Conference on Medical Image Computing and Computer Assisted Intervention*, 2015. 8
[37] Lingyan Ruan, Mojtaba Bemana, Hans-peter Seidel, Karol Myszkowski, and Bin Chen. Revisiting image deblurring with an efficient convnet. *arXiv preprint arXiv:2302.02234*, 2023. 8, 9, 10, 12
[38] Lingyan Ruan, Bin Chen, Jizhou Li, and Miuling Lam. Learning to deblur using light field generated and real defocus images. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 1, 2, 8, 9, 10
[39] Lingyan Ruan, Bin Chen, Jizhou Li, and Miu-Ling Lam. AIFNet: All-in-focus image restoration network using a light field-based dataset. *IEEE Transactions on Computational Imaging*, 7:675–688, 2021. 1, 9, 12
[40] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. 6
[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya

Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 8

[42] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Proc. of Asian Conference on Computer Vision*, 2018. 3

[43] Teresa Serrano-Gotarredona and Bernabé Linares-Barranco. A 128×128 1.5% contrast sensitivity 0.9% FPN 3 μs latency 4 mw asynchronous frame-free dynamic vision sensor using transimpedance preamplifiers. *IEEE Journal of Solid-State Circuits*, 48(3):827–838, 2013. 2

[44] Jianping Shi, Li Xu, and Jiaya Jia. Just noticeable defocus blur detection and estimation. In *Proc. of Computer Vision and Pattern Recognition*, 2015. 1, 2

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of International Conference on Learning Representations*, 2015. 8

[46] Irwin Sobel. An isotropic 3x3 image gradient operator. *Presentation at Stanford A.I. Project 1968*, 2014. 6

[47] Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *Proc. of International Conference on Computer Vision*, pages 2642–2650, 2021. 1, 2, 8, 9, 10

[48] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *Proc. of European Conference on Computer Vision*, 2022. 7

[49] Supasorn Suwajanakorn, Carlos Hernandez, and Steven M Seitz. Depth from focus with your mobile phone. In *Proc. of Computer Vision and Pattern Recognition*, 2015. 2, 5

[50] Minggui Teng, Chu Zhou, Hanyue Lou, and Boxin Shi. NEST: Neural event stack for event-based image enhancement. In *Proc. of European Conference on Computer Vision*, 2022. 2

[51] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time Lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 2

[52] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time Lens: Event-based video frame interpolation. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 2

[53] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proc. of International Conference on Computer Vision*, 2021. 3, 12

[54] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 8, 9, 10

[55] Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, and Tiejun Huang. Spk2imgnet: Learning to reconstruct dynamic scene from continuous spike stream. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 2, 3

[56] Junwei Zhao, Shiliang Zhang, Lei Ma, Zhaofei Yu, and Tiejun Huang. Spikingsim: A bio-inspired spiking simulator. In *IEEE International Symposium on Circuits and Systems*, 2022. 8, 9

[57] Wenda Zhao, Fei Wei, You He, and Huchuan Lu. United defocus blur detection and deblurring via adversarial promoting learning. In *Proc. of European Conference on Computer Vision*, 2022. 2, 8, 9, 10

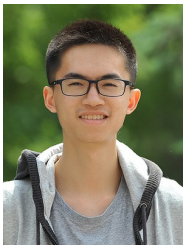
[58] Yajing Zheng, Lingxiao Zheng, Zhaofei Yu, Tiejun Huang, and Song Wang. Capture the moment: High-speed imaging with spiking cameras through short-term plasticity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8127–8142, 2023. 3

[59] Changyin Zhou, Daniel Miau, and Shree K. Nayar. Focal sweep camera for space-time refocusing. Technical report, Columbia University, 2012. 2, 3, 4, 5, 8, 11, 12

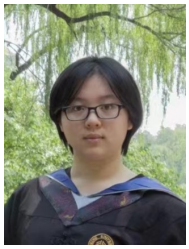
[60] Changyin Zhou and Shree Nayar. What are good apertures for defocus deblurring? In *Proc. of International Conference on Computational Photography*, 2009. 1

[61] 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*, 2019. 2, 3, 4, 6, 7, 8

[62] Lin Zhu, Xiao Wang, Yi Chang, Jianing Li, Tiejun Huang, and Yonghong Tian. Event-based video reconstruction via potential-assisted spiking neural network. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 3



Minggui Teng received the B.S. degree from Peking University, Beijing, China, in 2021. He is currently working toward the Ph.D. degree with the National Engineering Research Center of Video Technology, School of Computer Science, Peking University. His research interests are focused on neuromorphic camera and image enhancement. He has served as a reviewer for CVPR, ICCV, ECCV, *etc.*



Hanyue Lou received the B.S. degree summa cum laude from Peking University, Beijing, China, in 2023. She is currently working toward the Ph.D. degree with the National Engineering Research Center of Video Technology, School of Computer Science, Peking University. Her research interests are focused on applications of neuromorphic cameras.

Yixin Yang received her B.S. degree from Peking University, Beijing, China, in 2022. She is a Ph.D. candidate in the School of Computer Science, Peking University. Her research interests are focused on neuromorphic camera-based imaging and vision. She has served as a reviewer for CVPR, NeurIPS, *etc.*

Tiejun Huang (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science from Wuhan University of Technology, China in 1992 and 1995, respectively, and the Ph.D. degree in pattern recognition and image analysis from Huazhong (Central China) University of Science and Technology in 1998. He is currently a professor with the School of Computer Science, Peking University, and the Director of the Beijing Academy for Artificial Intelligence. His research areas include visual information processing and neuromorphic computing. He is a Fellow of CAAI, CCF, CSIG and vice chair of the China National General Group on AI Standardization. He published 300+ peer-reviewed papers on leading journals and conferences, and also co-editor of 4 ISO/IEC standards, 5 National standards and 4 IEEE standards. He holds 100+ granted patents. Professor Huang received National Award for Science and Technology of China (Tier-2) for three times (2010, 2012, 2017).

Boxin Shi (Senior Member, IEEE) received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the PhD degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Associate Professor (with tenure) and Research Professor at Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. His papers were awarded as Best Paper, Runners-Up at CVPR 2024, ICCP 2015 and selected as Best Paper candidate at ICCV 2015. He is an associate editor of TPAMI/IJCV and an area chair of CVPR/ICCV/ECCV. He is a senior member of IEEE.

Hybrid All-in-focus Imaging from Neuromorphic Focal Stack

Supplementary Material

Minggui Teng, Hanyue Lou, Yixin Yang, Tiejun Huang, *Senior Member, IEEE*, and Boxin Shi, *Senior Member, IEEE*

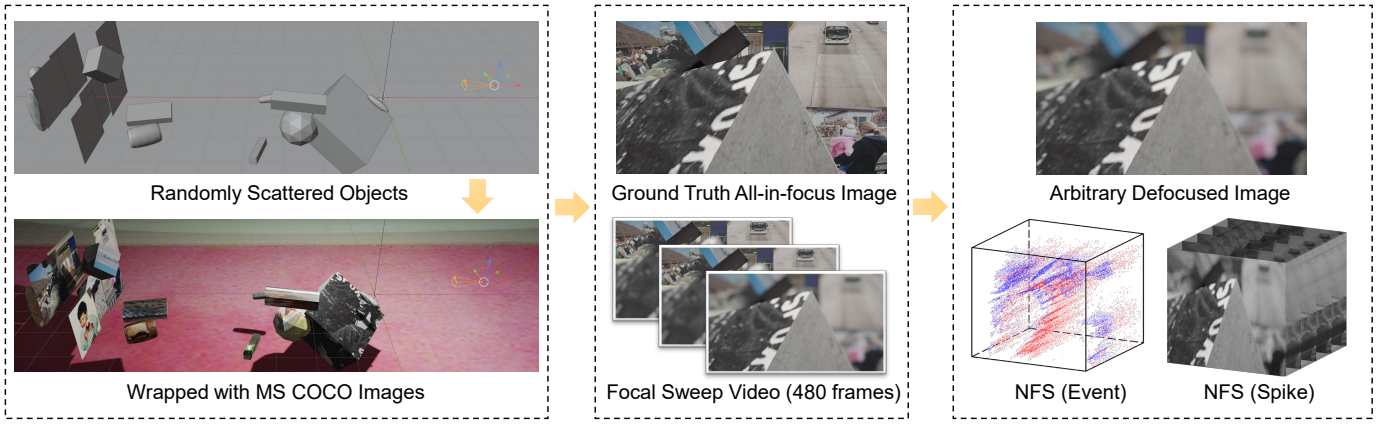


Figure 14: An illustration of our synthetic data generation pipeline. We simultaneous render NFS (event) and NFS (spike) with DVS-Voltmeter [5] and Spikesim [12] respectively.

6 Dataset

6.1 Synthetic dataset

The pipeline for generating the training dataset is shown in the Figure 14. First, we pick geometric objects, scale them, and scatter them at different depths in the scene randomly, forming 200 scenes. Second, to better match the data distribution to real-world images, we wrap the surfaces of the objects with images sampled from the MS-COCO dataset [6] as their textures. The sampled images are from *2017 Val images*, a subset of MS-COCO. The final step is to render 480 images as an image focal stack for each scene and then to render the ground truth image as an all-in-focus image with a small aperture.

After rendering the image focal stack, we input them into DVS-Voltmeter [5]/Spikesim [12] to generate event/spike streams. Since DVS-Voltmeter owns special parameters related to the camera type, to improve the generalization of the model to unknown types of event cameras, we apply the 6 different camera parameters in DVS-Voltmeter randomly. Each camera parameter ($k_1 \sim k_6$) is randomly sampled from the range $[Min, Max]$ shown in Table 3. The reference columns "DVS346" and "DVS240" are the parameters calibrated on the event camera models DVS346 and DVS240, provided as DVS-Voltmeter preset configurations [5].

Table 3: Settings of DVS-Voltmeter [5] parameters.

Param.	range	DVS346	DVS240
k_1	$[4.0, 5.5]$	5.3	4.4
k_2	$[18, 25]$	20	23
k_3	$[5 \times 10^{-5}, 2.5 \times 10^{-4}]$	1×10^{-4}	2×10^{-4}
k_4	$[0.8 \times 10^{-7}, 1.2 \times 10^{-7}]$	1×10^{-7}	1×10^{-7}
k_5	$[3 \times 10^{-9}, 8 \times 10^{-8}]$	5×10^{-9}	5×10^{-8}
k_6	$[8 \times 10^{-6}, 1.2 \times 10^{-5}]$	1×10^{-5}	1×10^{-5}

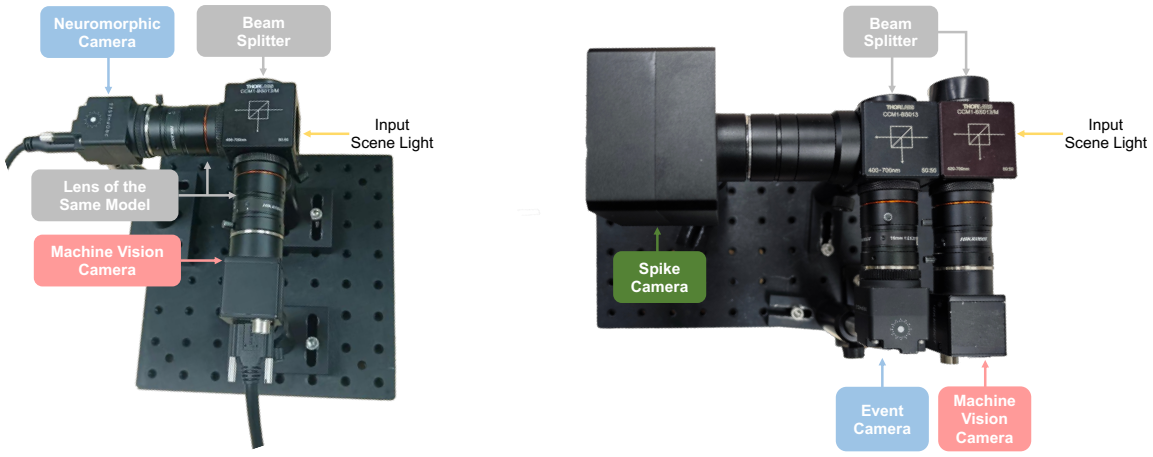
6.2 Real dataset

For real data acquisition, we design two types of hybrid camera systems. As illustrated in Figure 15, the system on the left consists of a dual-camera setup simultaneously capture images and event/spike data. And the right-side system features a three-camera setup capable of capturing images, event data, and spike data concurrently, providing comprehensive comparison of event and spike cameras.

7 Ablation Experiment

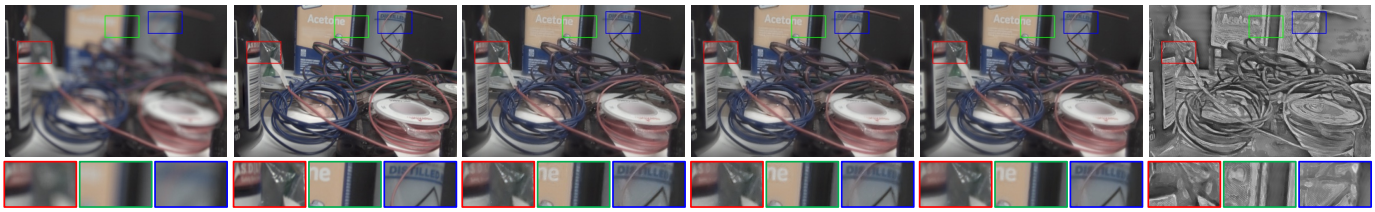
7.1 Loss function

We ablate different loss functions ($\mathcal{L}_2$ loss only, perceptual loss only, $\mathcal{L}_1$ loss + perceptual loss) from the complete model ($\mathcal{L}_2$ loss + perceptual loss) and evaluate them quantitatively in Table 4. Results show that the combination



Dual-Camera System Three-Camera System

Figure 15: An illustration of our hybrid camera systems. Left: The dual-camera system is designed for capturing images and event/spike data. Note that the neuromorphic camera is the event camera (PROPHESEE GEN4.0) in the figure for illustrative purposes, which can be substituted with a spike camera (Spike Camera-001T-Gen2). Right: The three-camera system captures images, event data, and spike data simultaneously for comprehensive event and spike comparison.



(a) Defocused Image (b) Ground Truth (c) Ours (d) Uniform (e) RNet+GDF (f) ET+MNet

Figure 16: Visual quality comparison of ablation studies on synthetic data.

Table 4: Ablation study on loss functions.

	PSNR ↑	SSIM ↑	MS-SSIM ↑	LPIPS ↓
$\mathcal{L}_2$ only	33.04	0.9305	0.9605	0.1547
perc. only	31.51	0.8977	0.9567	0.1580
$\mathcal{L}_1$ + perc.	33.09	0.9300	0.9599	0.1515
$\mathcal{L}_2$ + perc.	33.25	0.9323	0.9611	0.1510

of $\mathcal{L}_2$ loss and perceptual loss improves the performance of NeuroRefocusNet and NeuroMergeNet in reconstructing all-in-focus images. Given that our network architecture is consistent for both event and spike cameras, ablation studies are conducted on one type of neuromorphic camera.

7.2 Qualitative comparison

The qualitative comparison among the different ablation studies is shown in Figure 16. According to the results, our complete model can produce a sharper, all-in-focus image. Note that ET-Net [10] only reconstructs gray-scale images, thus, we only compare the results of "ET+MNet" with the gray-scale ground truth.

7.3 Analysis

To verify the effectiveness of each module, we conduct three ablation studies, shown Section 4.4, and the detailed analysis of each ablation study is listed as follows :

- "ET+MNet": Our NeuroRefocusNet takes a single defocused image with the corresponding event stream as input, while ET-Net [10] only utilizes the event stream, resulting in a lack of texture details.
- "RNet+GDF": Since the event stream provides high-temporal-resolution edge information, compared with gradient domain fusion [14], our NeuroMergeNet can predict more accurate weights for focal stack merging.
- "Uniform": By dynamically selecting refocus distances with our golden search method instead of sampling distances uniformly, our method can refocus to objects that would fall between the uniform samples otherwise, as illustrated in Figure. 2 (the blue cube is out of focus in all focal stack images). Our method also avoids refocusing on distances with no objects, which causes a waste of computation when distances are sampled uniformly.

8 Input Quantities of Focal Stack

In this section, we conduct an experiment with focal stacks containing different numbers of images. The experiments involve specifically designed sets with quantities of

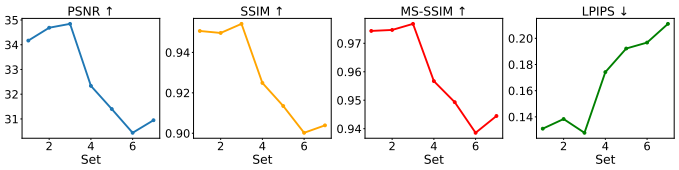


Figure 17: Quantitative comparison of different image quantities in the focal stacks on the LiFF dataset [1]. The sets 1 to 6 contain image quantities ranging from $\{30, 25, 20, 15, 10, 8, 6\}$, respectively.

Table 5: Quantitative results on the LiFF dataset [1].

	PSNR ↑	SSIM ↑	MS-SSIM ↑	LPIPS ↓
KPAC [9]	27.96	0.8396	0.9115	0.2473
IFAN [4]	29.59	0.8119	0.8679	0.3741
APL [13]	27.14	0.7758	0.8522	0.4060
DRBNet [8]	30.51	0.8639	0.9278	0.2164
Ours (ev)	33.25	0.9323	0.9611	0.1510
Ours (sp)	34.85	0.9541	0.9768	0.1278

$\{30, 25, 20, 15, 10, 8, 6\}$, corresponding to sets 1 to 6, respectively. The results of this experiment are shown in Figure 17. As the number of images in the focal stack increases, there is a decrease in the quality of the all-in-focus images produced. Furthermore, our method identifies the optimal quantity of the image focal stack beyond which the quality of the output diminishes.

9 Experiments with Image-based Methods

To compare with single-image-based methods comprehensively, we feed them with 10 images in the same scene, which are focused at different focal distances, obtain the 10 defocused deblurring images, and then calculate the average metric values as the final results. The quantitative result is shown in Table 5. Based on the results, our method still outperforms the state-of-the-art image-based methods.

10 Speed Variation Issue of Focal Plane

For convention image focal stack methods [2], [14], the focal plane must move at a stable speed. However, our NFS is less affected by this restriction. We take NFS by rotating the focus ring by hand, which inevitably makes the focal plane move at a varying speed. With the high temporal resolution property, the neuromorphic camera can detect scene radiance changes at the microsecond level. Since our manual rotation speed is much slower than its temporal resolution, the performance of our method is robust to such speed variation. Since the quantitative results have shown spike cameras have better reconstruction ability, we choose events camera for comparison in this scenario without loss of generality. As the example in Figure 18 shows, our method can restore an all-in-focus image with consistently high quality, given NFS captured at different speeds. As we capture NFS manually, we show histograms of the number of events at each timestamp to partially reflect the speed variation when rotating the focus ring for NFS capture.

11 Network Details

In this section, we present architecture details of our NeuroRefocusNet (shown in Table 6) and NeuroMergeNet (shown in Table 7).

12 Image Focal Stack

Our method can restore images refocused at arbitrary focus distances from a single defocused image and the corresponding NFS. The generated image focal stacks are shown in our supplementary video.

13 More Results on Synthetic Dataset

In this section, we provide more qualitative comparisons among our method (ev/sp), DRBNet [8], IFAN [4], KPAC [9], APL [13], Restormer [11], and LaKDNet [7] on synthetic data, shown in Figure 19, and Figure 20.

14 More Results on Real Dataset

In this section, we provide more qualitative comparisons among our method (ev/sp), DRBNet [8], IFAN [4], KPAC [9], APL [13], Restormer [11], and LaKDNet [7] on real data, shown in Figure 21 and Figure 22.

References

[1] Donald G. Dansereau, Bernd Girod, and Gordon Wetzstein. LiFF: Light field features in scale and depth. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 3
[2] Berthold Klaus Paul Horn. Focusing. Technical report, MIT, 1968. 3
[3] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 4
[4] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 3, 5, 6, 7, 8
[5] Songnan Lin, Ye Ma, Zhenhua Guo, and Bihan Wen. DVS-Voltmeter: Stochastic process-based event simulator for dynamic vision sensors. In *Proc. of European Conference on Computer Vision*, 2022. 1
[6] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. of European Conference on Computer Vision*, 2014. 1
[7] Lingyan Ruan, Mojtaba Bemana, Hans-peter Seidel, Karol Myszkowski, and Bin Chen. Revisiting image deblurring with an efficient convnet. *arXiv preprint arXiv:2302.02234*, 2023. 3, 5, 6, 7, 8
[8] Lingyan Ruan, Bin Chen, Jizhou Li, and Miuling Lam. Learning to deblur using light field generated and real defocus images. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 3, 5, 6, 7, 8
[9] Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *Proc. of International Conference on Computer Vision*, pages 2642–2650, 2021. 3, 5, 6, 7, 8
[10] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proc. of International Conference on Computer Vision*, 2021. 2
[11] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 3, 5, 6, 7, 8
[12] Junwei Zhao, Shiliang Zhang, Lei Ma, Zhaofei Yu, and Tiejun Huang. Spikingsim: A bio-inspired spiking simulator. In *IEEE International Symposium on Circuits and Systems*, 2022. 1
[13] Wenda Zhao, Fei Wei, You He, and Huchuan Lu. United defocus blur detection and deblurring via adversarial promoting learning. In *Proc. of European Conference on Computer Vision*, 2022. 3, 5, 6, 7, 8
[14] Changyin Zhou, Daniel Miau, and Shree K. Nayar. Focal sweep camera for space-time refocusing. Technical report, Columbia University, 2012. 2, 3

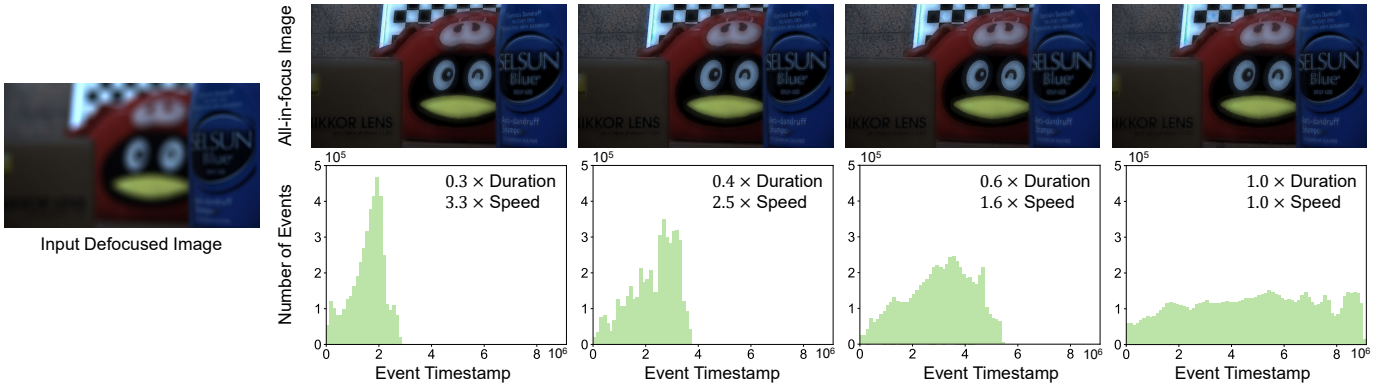


Figure 18: Results of our method given NFS (event) captured at different speeds. Below each image, we show the histograms of the number of events at each timestamp, to partially reflect the speed variation when rotating the focus ring for NFS (event) capture.

Table 6: Network details of NeuroRefocusNet. DenseConv modules are densely connected convolutional blocks [3]. ResBlock modules are residual blocks. Deconv modules are transposed convolutional blocks. All modules include batch normalization and activation functions.

NeuroRefocusNet	Input	Kernel Size	Stride	In Channels	Out Channels	Output
Conv1	Input Image	3	1	3	64	conv1
DenseConv1	conv1	3	1	64	128	denseconv1
Conv2	denseconv1	2	2	128	128	conv2
DenseConv2	conv2	3	1	128	256	denseconv2
Conv3	denseconv2	2	2	256	256	conv3
DenseConv3	conv3	3	1	256	512	denseconv3
ConvN1	NFS	3	1	64	64	convn1
DenseConvN1	convn1	3	1	64	128	denseconvn1
ConvN2	denseconvn1	2	2	128	128	convn2
DenseConvN2	convn2	3	1	128	256	denseconvn2
ConvN3	denseconvn2	2	2	256	256	convn3
DenseConvN3	convn3	3	1	256	512	denseconvn3
Deconv2	[denseconv3, denseconvn3]	2	2	1024	256	deconv2
Conv5	[denseconv2, denseconvn2, deconv2]	1	1	768	128	conv5
DenseConv5	conv5	3	1	128	256	denseconv5
Deconv1	denseconv5	2	2	256	128	deconv1
Conv6	[denseconv1, denseconvn1, deconv1]	1	1	384	64	conv6
ResBlock1	conv6	3	1	64	64	resblock1
ResBlock2	resblock1	3	1	64	64	resblock2
PredConv	resblock2	3	1	64	3	pred

Table 7: Network details of NeuroMergeNet. DenseConv modules are densely connected convolutional blocks [3]. ResBlock modules are residual blocks. Deconv modules are transposed convolution blocks. All modules include batch normalization and activation functions.

NeuroMergeNet	Input	Kernel Size	Stride	In Channels	Out Channels	Output
Conv1	One of Image Stack + NFS	3	1	4	64	conv1
DenseConv1	conv1	3	1	64	128	denseconv1
Conv2	denseconv1	2	2	128	128	conv2
DenseConv2	conv2	3	1	128	256	denseconv2
Conv3	denseconv2	2	2	256	256	conv3
DenseConv3	conv3	3	1	256	512	denseconv3
Deconv2	denseconv3	2	2	512	256	deconv2
Conv5	[denseconv2, deconv2]	1	1	512	128	conv5
DenseConv5	conv5	3	1	128	256	denseconv5
Deconv1	denseconv5	2	2	256	128	deconv1
Conv6	[denseconv1, deconv1]	1	1	256	64	conv6
ResBlock1	conv6	3	1	64	64	resblock1
ResBlock2	resblock1	3	1	64	1	intial weights

Figure 19: Visual quality comparison with image-based defocus deblurring methods on synthetic data (Part I). Visual quality comparison with image-based defocus deblurring methods on real data. (a) Defocused image. (b) Events. (c) Spikes. (d)∼(k) All-in-focus results of DRBNet [8], IFAN [4], KPAC [9], APL [13], Restormer [11], LaKDNet [7], ours (ev), and ours(sp). (l) Ground Truth.

Figure 20: Visual quality comparison with image-based defocus deblurring methods on synthetic data (Part II). Visual quality comparison with image-based defocus deblurring methods on real data. (a) Defocused image. (b) Events. (c) Spikes. (d)∼(k) All-in-focus results of DRBNet [8], IFAN [4], KPAC [9], APL [13], Restormer [11], LaKDNet [7], ours (ev), and ours(sp). (l) Ground Truth.

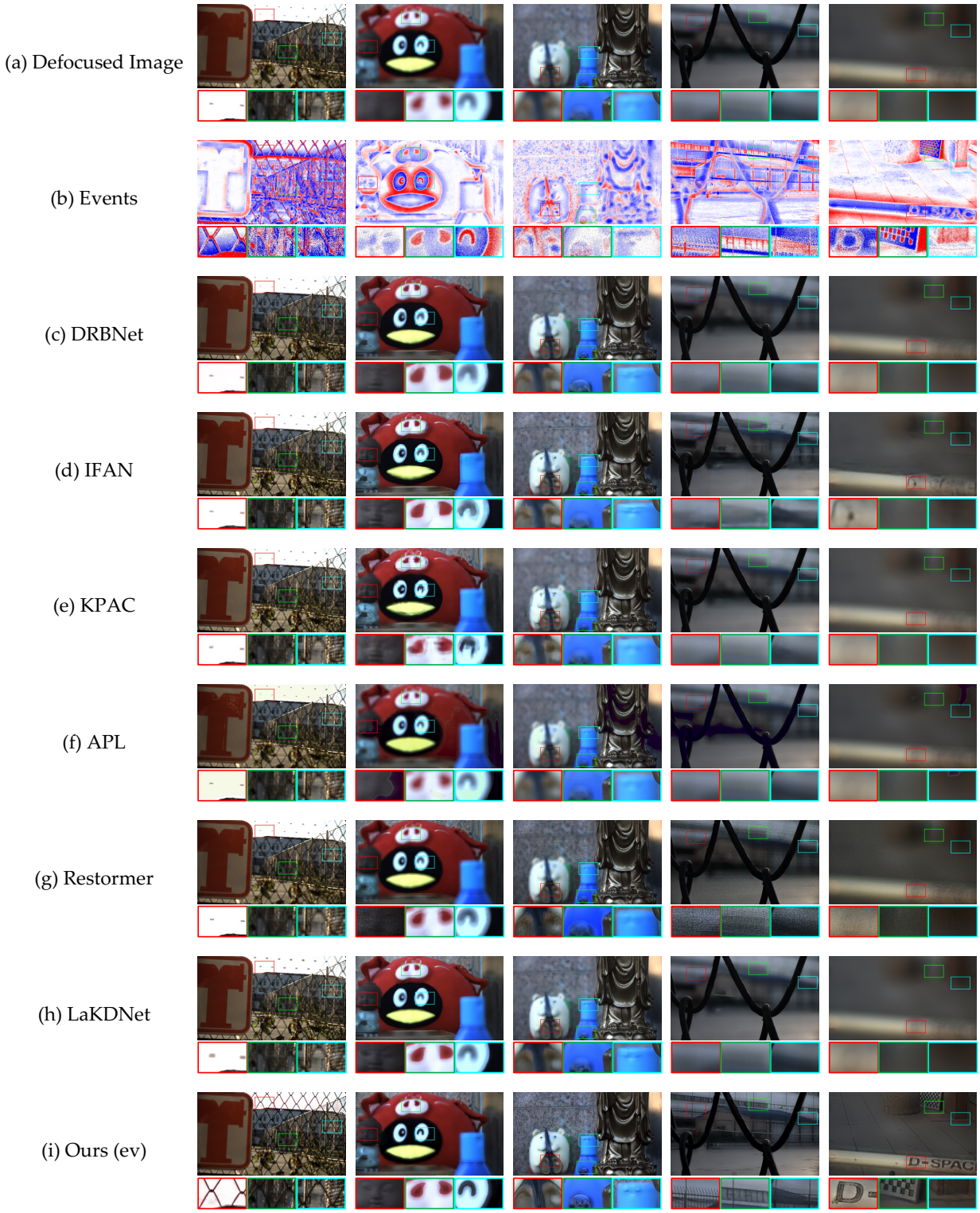


Figure 21: Visual quality comparison with image-based defocus deblurring methods on real data (Part I). (a) Defocused image. (b) Events. (c)∼(i) All-in-focus results of DRBNet [8], IFAN [4], KPAC [9], APL [13], Restormer [11], LaKDNet [7], and ours (ev).

Figure 22: Visual quality comparison with image-based defocus deblurring methods on real data (Part II). (a) Defocused image. (b) Spikes. (c)∼(i) All-in-focus results of DRBNet [8], IFAN [4], KPAC [9], APL [13], Restormer [11], LaKDNet [7], and ours (sp).