Coded Event Focal Stack for Continuous Refocusing in Dynamic Scene

Minggui Teng†, Suhang Xuan†, Zhiang Yan, Hanyue Lou, Boyu Li,
Bin Fan, Boxin Shi‡, *Senior Member, IEEE*

***Abstract*—Traditional cameras face limitations in maintaining focus across dynamic scenes, especially during rapid motion, due to the constraints of their lenses. Post-capture refocusing techniques, including deep learning-based methods and light field cameras, have been explored to mitigate these challenges. However, these approaches frequently struggle with temporal consistency or experience a trade-off in spatial resolution. In this paper, we introduce the coded event focal stack, a novel approach that captures both motion and depth information through event streams recorded during a modulated focal sweep. Our coded event focal stack enables the generation of full-time intermediate frames refocused at arbitrary focal distances. Extensive experiments on both synthetic and real-world datasets demonstrate the superior refocusing capability of our method over state-of-the-art techniques, particularly in dynamic scenes with complex motion and depth variations.**

***Index Terms*—Computational Photography, Event Camera, Image Refocusing.**

I. Introduction

Due to the characteristics of lens module design, only objects located within the depth of field (DoF) appear sharp and clear in images. Specifically, when the aperture is set to a smaller size, the DoF increases accordingly, rendering more of the scene in sharp focus. However, this setting also has the side effects of prolonging the exposure time and necessitating a higher ISO value. Conversely, using a larger aperture narrows the DoF but allows for shorter exposure times. In practice, a larger aperture is frequently required to prevent motion blur for dynamic scenes, yet this simultaneously intensifies the challenge of focus tracking.

For focus tracking, traditional cameras employ phase detection [4], [51] or contrast maximization [15], [21], [63] techniques. Phase detection struggles in low-light conditions, while contrast maximization requires multiple iterations to focus on an object properly. When an object moves rapidly or the lighting conditions are complex, traditional autofocus systems often fail to maintain focus tracking. Thus, the post-capture refocusing algorithm (*e.g.*, light-field camera refocusing [35]) is essential in the dynamic scene. With advancements in deep learning, it is now possible to refocus a single object by combining defocus deblurring methods [26], [43], [44], [52] and adding defocus blur [16], [38], [62]. RefocusGAN [46] is one of the first attempts to perform image refocusing from a single image by leveraging the generative capabilities of generative adversarial networks (GAN) [14] via generating all-in-focus imaging before restoring refocused images. However, these two-step methods still struggle to maintain temporal consistency in dynamic scenes and may risk losing high-frequency details (shown in the top row of Figure 1 (c)).

In dynamic scenes, video refocusing methods [22], [41] rely on inter-frame differences to estimate depth maps, which are then used for image defocus deblurring and refocusing. However, these methods inherently involve image deconvolution, resulting in unnatural refocused outcomes. With the advent of specialized deformable lenses, frame-based focal stack methods [34] can capture focal stacks at approximately 4 Hz, cooperating with 120 frames per second camera, allowing for video refocusing by selecting frames where the focal plane matches object depths. Nonetheless, the limited frame rate of traditional cameras poses challenges in capturing continuous depth and motion information, thereby reducing the effectiveness of refocusing in dynamic scenes.

In recent years, event cameras [8], [48] have demonstrated high temporal resolution in recording dynamic information, introducing new possibilities for all-in-focus imaging by capturing an event focal stack [33], [53]. This approach leverages the high temporal resolution of event cameras to accurately and rapidly capture intensity variations during focal sweeps. However, current methods [33], [53] are predominantly restricted to intensity changes in static scenes and cannot effectively handle dynamic scenes, as illustrated in the middle row of Figure 1(c). In contrast, light field with event camera [39] capturing methods require more complex hardware setups and incur reduced spatial resolution compared to focal stack approaches. Thus, the application of event-based focal stack techniques in dynamic scenes remains relatively unexplored. Given the microsecond-level sensitivity of event cameras to dynamic information, we are inspired to ask: *Can we generate full-time intermediate frames refocused at arbitrary focal distances in dynamic scenes by leveraging event focal stacks?*

In this paper, we introduce the *Coded Event Focal Stack* (CEFS), a novel technique that leverages event streams captured during a focal sweep, which is modulated with a designated cycle to record variations caused by both object motion and the focal sweep (shown in the bottom row of Figure 1 (b)). Building on the observation that events generated by focal sweeps exhibit a symmetric distribution, we propose to effectively decompose motion and depth information from CEFS.

† Contributed equally to this work as first authors.
‡ Corresponding author: shiboxin@pku.edu.cn
Minggui Teng, Suhang Xuan, Zhiang Yan, Hanyue Lou, Boyu Li and Boxin Shi are with State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing 100871, China.
Bin Fan is with School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China

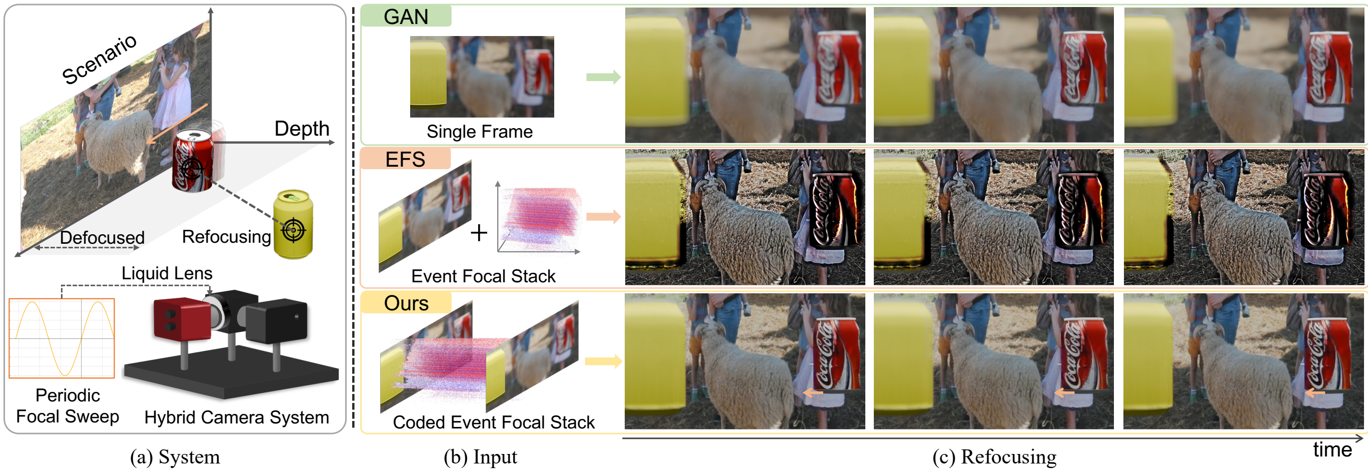


Fig. 1. (a) We design a hybrid camera system for dynamic scene refocusing, utilizing a periodic event focal sweep modulated by a liquid lens. Given the images focused on the foreground yellow can, the goal is to refocus frames onto the fast-moving red can. (b) This goal can be achieved by using a single image (RefocusGAN [46], top row) or additionally capturing the event focal stack (EFS [33], middle row). (c) Our approach (bottom row) produces sharp and temporally consistent results, while RefocusGAN [46] cannot recover high-frequency details and EFS [33] fails to handle motion. Note that only our approach can produce a full-time sequence of intermediate frames, while other methods must rely on an additional video frame interpolation model (*e.g.*, RIFE [19]) to generate intermediate frames.

Using the motion information within CEFS, we first warp two consecutive frames into specific timestamps to restore the intermediate frame, after which the focal sweep information allows us to refocus this intermediate frame into any desired focal distance. To determine the appropriate refocusing focal distance, we introduce an iterative search scheme to identify the start and end points of the focal distance range. As a result, our approach enables the generation of full-time intermediate frames refocused to arbitrary distances in the dynamic scene (shown in the bottom row of Figure 1 (c)). Our contributions are demonstrated by proposing:

- a novel focal stack capturing technique, termed coded event focal stack, with modulated periodic event focal sweep, continuously recording variations caused by motion and focal distance change;
- a two-stage framework decomposing depth and motion information for intermediate frame generation and image refocusing, based on the symmetric distribution prior; and
- an iterative search strategy on the selective region of interest for post-capturing focus tracking.

We construct a hybrid camera system and modulate the focal sweep up to 50 Hz using a liquid lens to validate our method in real-world scenarios (Figure 1 (a)). The experiments demonstrate that our approach can obtain superior quality in refocusing images over state-of-the-art methods in both synthetic and real datasets.[1]

II. Related Work

Most refocusing methods consist of two steps: defocus deblurring and refocusing. We first review defocus deblurring methods, covering both image-based and computational photography approaches. Then, we review refocusing methods based on a single image or video, alongside event camera guided image restoration.

[1]The video results can be found in the supplementary material.

Defocus deblurring. Image defocus deblurring [13], [20], [25], [37], [49], [66] can be divided into two steps: they first estimate the defocus map, and then apply the non-blind deconvolution. The quality of results highly depends on the accuracy of the defocus map. With deep network developing, end-to-end methods [42], [44], [52] have demonstrated superior performance. However, it is difficult to recover high-frequency regions from a defocused image, and these methods do not generalize well in real-world scenarios. Computational photography-based defocus deblurring methods typically employ specialized image capture pipelines to obtain additional information from different focal planes, thereby enabling more effective restoration of defocused regions. Representative approaches include focal stack methods [17], [67] and focal sweep techniques [24]. Alternatively, some methods leverage unconventional optical elements designed to modify blur characteristics and alleviate the ill-posedness of defocus deblurring. These elements include coded apertures [27], which modify blur characteristics in the frequency domain to facilitate restoration; wavefront coding [11], which alters the phase of incoherent wavefronts to maximize optical power at the imaging plane; lattice lenses [28], optimized to enhance the defocus spectrum in frequency-domain analysis; and dual-pixel sensors [1], [2], which utilize two sub-aperture images to effectively mitigate defocus blur. In addition to modifying optical components, another line of work [29] focuses on adapting the capture methodology itself, rather than altering the lens, to improve the deconvolution process and reduce the complexity of the required optical setup.

Image/video refocusing. RefocusGAN [46] is one of the first single image based refocusing methods, which utilizes GAN model [14] to recover an AIF image and refocus this image into the desired focal distance. Alzayer *et al.* [3] propose a dual-camera system enabling refocusing by fusing images from both a wide field of view (FoV) and an ultra-wide FoV.

TABLE I
A SUMMARY OF DYNAMIC REFOCUSING METHODS. FOR FOCAL SWEEP BASED METHODS [22], [34], "FREQUENCY" INDICATES THE NUMBER OF FOCAL SWEEPS PER SECOND. "FOCAL DISTANCE" REPRESENTS THE FOCAL DISTANCE RANGE OF REFOCUSED FRAMES. "TIME" DESCRIBES WHETHER THIS METHOD CAN GENERATE A CONTINUOUS SEQUENCE OF INTERMEDIATE FRAMES.

Method	Frequency	Capturing Device	Inputs	Focal Distance	Time
Ruan *et al*. [41]	–	Traditional camera	Defocused frames	Fore/Background	N
Miau *et al*. [34]	4.3 Hz	Deformable lens	Sharp focal stacks	Fixed depth	N
Kim *et al*. [22]	4 Hz	Frame camera + Manual rotation	Focal sweep videos	Arbitrary	N
Ours	50 Hz	RGB camera / Event camera + Liquid lens	CEFS + Arbitrary-focused frames	Arbitrary	Y

Light field cameras capture both the intensity and direction of light rays, allowing for post-capture refocusing within the scene [30], [35]. However, these methods have difficulty in processing temporally varying defocus blur.

To handle temporal variations, Ruan *et al*. [41] proposed using neural networks to implicitly represent scene information and extract refocused images through atlas learning. Miau *et al*. [34] designed a deformable lens to capture image focal stacks with a fixed frequency. Selecting frames where the focal plane matches the object depths enables focus tracking in video. Kim *et al*. [22] proposed capturing focal sweeping videos, deconvolving defocused frames to generate RGB-D frames, and then refocusing based on the estimated depth map. As shown in Table I, these dynamic refocusing methods have difficulty refocusing images to the arbitrary focal distance at the intermediate timestamp.

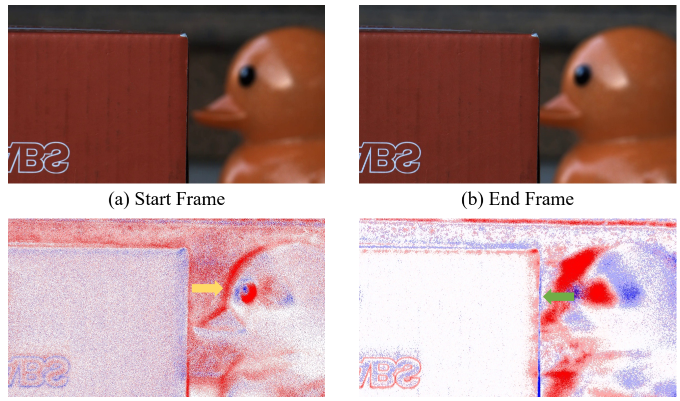
Event-guided image restoration. Event camera owns microsecond-level sensitivity, thus, it has been applied for high frame rate video generation [40], [60], [61], image deblurring [7], [36], [54], [65], [68], and image interpolation [55], [56]. Recently, Lou *et al*. [33] proposed an event focal stack (EFS), which encodes continuous depth information by EFS in static scenes. Guided by the texture information inside EFS, they reconstructed a refocused image to form an image focal stack. However, this method is not suitable for dynamic scenes, as they did not consider the events triggered by object motion or camera shake. Inspired by the ability of event streams to capture continuous intensity changes caused by both focal sweep and object motion, this paper investigates the utilization of periodic focal sweeps with event cameras to overcome the limitations of existing refocusing methods in dynamic scenes.

III. PROPOSED METHOD

In this section, we first introduce the preliminary concepts in Section III-A. Subsequently, we formulate the coded event focal stack and introduce our model for reconstructing refocused images in Section III-B. Guided by the coded event focal stack, we propose an event-guided dynamic refocusing framework detailed in Section III-C. Furthermore, we present a focus tracking strategy in Section III-D. Our dataset synthesis method is described in Section III-E, and implementation details are elaborated in Section III-F.

A. Preliminaries

Event formulation. Event cameras [48] are bio-inspired cameras that record scene radiance change with discrete event



(a) Start Frame (b) End Frame (c) Sweep-triggered Events (d) Motion-triggered Events

Fig. 2. An example of CEFS visualization. (a)/(b) Two consecutive frames focused at the same distance. (c)/(d) Visualizations of events triggered by focal sweep and object motion, respectively. The yellow arrow indicates the shifting focal distance from the box to the duck, where sweep-triggered events are distributed across the entire image. The green arrow denotes the motion direction, with motion-triggered events aligned along the object's trajectory. Notably, sweep-triggered events are concentrated around the duck's boundary with predominantly positive polarity (red), suggesting that the image is gradually refocusing on the duck.

streams. An event signal $e = (\mathbf{x}, p, t)$ is triggered when the intensity change exceeds the preset threshold c at pixel $\mathbf{x}$, *i.e.*,

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

in which $\mathbf{I}_{\mathbf{x}}^{t}$ and $\mathbf{I}_{\mathbf{x}}^{t-\Delta t}$ represent the intensity radiance at time t and $t - \Delta t$, respectively. Polarity $p \in \{1, -1\}$ indicates intensity change is increase or decrease. Since Equation (1) applies to each pixel $\mathbf{x}$ independently, pixel indices are omitted henceforth.

Given two consecutive frames $\mathbf{I}^{t_0}$ and $\mathbf{I}^{t_1}$, there are N_e events that occur between these two frames. Since the event streams record intensity changes between them, we can hence link them with event integration, *i.e.*,

$$\mathbf{I}^{t_1} = \mathbf{I}^{t_0} \exp\left(\sum^{N_e} c_t \cdot e(t)\right), \tag{2}$$

where $e(t),\ t \in [t_0, t_1]$ represents the event triggered between two frames, and c_t is the spatial-temporal variant threshold [18].

Defocus formulation. Image defocus is formulated as a convolution with a depth-variant point spread function (PSF).

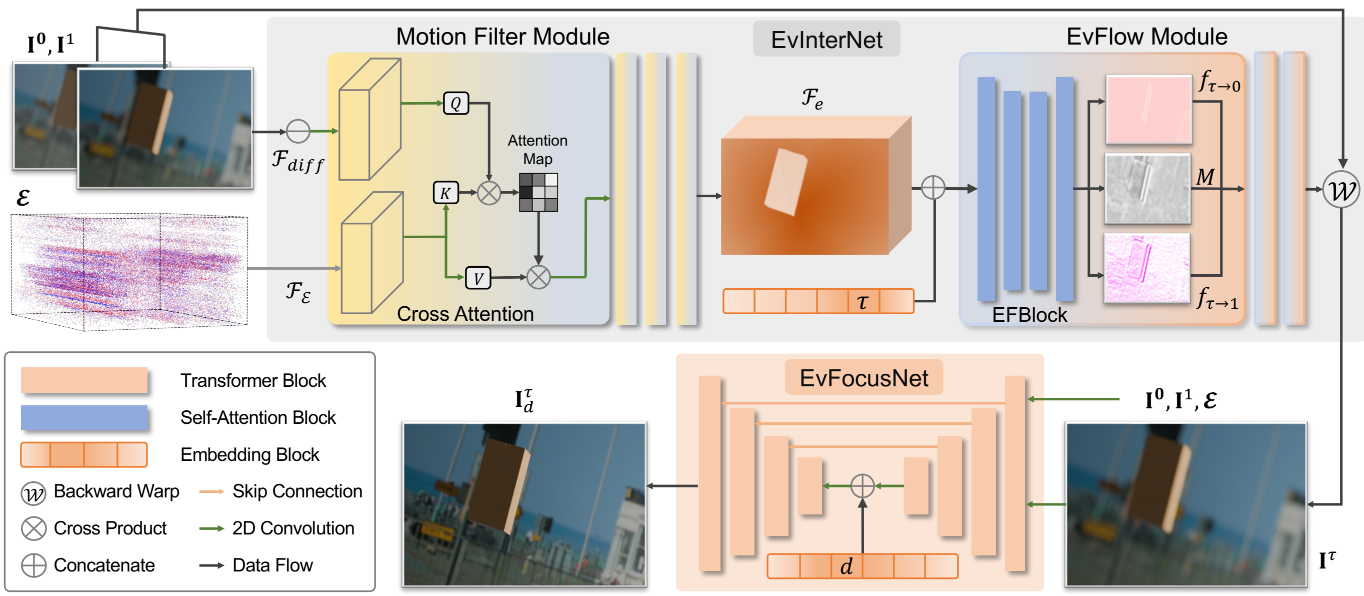


Fig. 3. Pipeline of our proposed event-guided dynamic refocusing framework. Given two boundary frames $\mathbf{I}^0$ and $\mathbf{I}^1$ together with the CEFS stream $\mathcal{E}$, the framework synthesizes a refocused intermediate frame $\mathbf{I}_d^\tau$ at any time τ and target depth d. The system first extracts motion-triggered events using the Motion Filter, a cross-attention module that leverages the frame-difference feature $\mathcal{F}_{\texttt{diff}}$ to suppress sweep-triggered responses and isolate motion-relevant signals, producing purified event features $\mathcal{F}e$. These features are then used by EvFlow, which predicts bi-directional optical flow $(f_{\tau\to 0}, f_{\tau\to 1})$ and an adaptive fusion mask M to warp and blend the boundary frames into a temporally coherent intermediate frame $\mathbf{I}^\tau$. Finally, EvFocusNet fuses $\mathbf{I}^\tau$ with CEFS and boundary textures through transformer-based depth-conditioned refinement, where the focal distance d is encoded to guide texture alignment and sweep-based depth inference, producing the final refocused output $\mathbf{I}_d^\tau$.

The PSF of defocus blur is a Circle of Confusion (CoC) and is often modeled as a Gaussian function [24]:

$$\mathcal{P}(r, u, d) = \frac{2}{\pi(g\mathcal{D}(u-d))^2} \exp\left(-\frac{2r^2}{(g\mathcal{D}(u-d))^2}\right), \quad (3)$$

where r represents the distance of an image point from the center of the PSF, u denotes the ideal focal distance, d is the object focal distance, $\mathcal{D}(u-d)$ is the diameter of the CoC related to the difference of u and d. The ideal focal distance u is derived from the Thin Lens Law: $1/f = 1/u + 1/v$. Specifically, when $d = u$ (*i.e.*, ideally in-focus), $\mathcal{P}(r, u, d)$ becomes an impulse function.

B. Coded event focal stack

By applying the focal sweep technique to an event camera [33], [53], continuous event streams can be produced, which capture changes in pixel radiance as the focal plane sweeps across the scene. Combining Equation (1) and Equation (3), we can derive:

$$\log\left(\frac{\mathcal{P}(r, u+\Delta u, d) * \mathbf{I}}{\mathcal{P}(r, u, d) * \mathbf{I}}\right) = p \cdot c, \quad (4)$$

where $\mathbf{I}$ is the in-focused object radiance value and $*$ denotes the convolution operator. According to Equation (4), during the focal sweep, the change of PSF will trigger events, thus the static region will also trigger events. On the other hand, the events triggered by object motion are located perpendicular to the direction of motion [12]. As shown in Figure 2, the events triggered by focal sweep are uniformly distributed in the scene, while the events triggered by object motion are alongside the motion trajectory.

Given two consecutive frames, $\mathbf{I}^0$ and $\mathbf{I}^1$, both focused at the focal distance d_0, we can establish a link to these two frames using event streams. To achieve this, we conduct a focal sweep with event cameras, where the focal distance starts from d_0 and shifts to $d_0 + \Delta d$ within a frame interval, *i.e.*, $\Delta d = f(t), t \in [t_0, t_1]$. Here, $f(t)$ is a function of t that describes how the focal depth changes over time. We refer to recorded event streams as Coded Event Focal Stack (CEFS), as we modulate the start and end position of the focal sweeps, with time during the interval of the consecutive frames. CEFS consists of two types of events: *motion-triggered* events $\mathcal{E}_p$, which are passively generated by scene motion, and *sweep-triggered* events $\mathcal{E}_a$, which are actively generated by the focal sweep. *i.e.*,

$$\begin{aligned} \mathcal{E} &= \{e_p\}_{t_0}^{t_1} \cup \{e_a\}_{d_0}^{d_0+\Delta d} \\ &= \mathcal{E}_p \cup \mathcal{E}_a. \end{aligned} \quad (5)$$

In our setup, $f(t)$ is the sine function with a period of $t_1 - t_0$.

Once CEFS can be separated into motion-triggered and sweep-triggered types, we can generate frame $\mathbf{I}_d^\tau$ refocused at focal distance d on intermediate time τ between $[t_0, t_1]$ with Equation (2):

$$\mathbf{I}_d^\tau = \mathcal{C}\left(\mathbf{I}^0 \cdot \mathcal{I}(\mathcal{E}_p^{t_0\to\tau}), \mathbf{I}^1 \cdot \mathcal{I}(\mathcal{E}_p^{t_1\to\tau})\right) \cdot \mathcal{I}(\mathcal{E}_a^{d_0\to d}), \quad (6)$$

where $\mathcal{I}(\cdot)$ is the integral operator, and $\mathcal{C}(\cdot)$ is the merge operator. Therefore, the core step is the separating strategy, dividing events from its triggering source, and then we can apply them with Equation (6) to refocus frames in the dynamic scene.

Event separation. As shown in Equation (6), the core challenge lies in separating the two types of events. According to Equation (4), periodic focal sweeping causes the point spread function (PSF) to vary periodically, generating sweep-triggered events that are temporally symmetric within a complete sweep cycle [5]. When we integrate all event signals e in the CEFS $\mathcal{E}$ over an entire sweep, this temporal symmetry causes sweep-triggered events to cancel out, so that the aggregated signal effectively approximates the sum of motion-triggered events, with the contribution of sweep-triggered events approaching zero,[2], *i.e.*,

$$\sum\nolimits_{e\in\mathcal{E}} p\cdot c \approx \sum\nolimits_{e\in\mathcal{E}_p} p\cdot c, \tag{7}$$

where polarity p is a binary value (1 or -1). By combining Equation (2) and Equation (7), the start frame $\mathbf{I}^0$ and the end frame $\mathbf{I}^1$ can be warped to the target intermediate frame, which is focused at the same depth. In practice, although the thickness modulation of a real liquid lens is not perfectly periodic and may exhibit small delays, this does not undermine our method: because the event-triggering mechanism depends only on changes in depth rather than on the exact sweep speed, it is sufficient that the lens returns to the same focal position at the end of each sweep cycle.

Due to the presence of noise signals in practice, directly applying Equation (2) and Equation (7) introduces artifacts. Since $\mathbf{I}^0$ and $\mathbf{I}^1$ are focused at the same depth, their differences are primarily caused by motion, which can also be represented by motion-triggered events, *i.e.*,

$$\mathbf{I}^{\texttt{diff}} = \log\mathbf{I}^1 - \log\mathbf{I}^0 \approx \sum\nolimits_{e\in\mathcal{E}_p} p\cdot c. \tag{8}$$

$\mathbf{I}^{\texttt{diff}}$ can serve as a cue to extract motion-triggered events from the CEFS, thereby mitigating the impact of noise-induced artifacts. Moreover, the threshold of the event camera is not constant in real-world scenarios [18], which poses additional challenges for refocusing frames from CEFS. To tackle these challenges, we introduce a CEFS-guided frame refocusing framework that leverages event characteristics to achieve more robust and accurate refocusing, as detailed below.

C. Dynamic refocusing from CEFS

The overall framework of our method is illustrated in Figure 3. To address the challenges of camera noise and non-constant threshold, we propose EvInterNet, which extracts motion-triggered events from the CEFS by effectively fusing both local and global information. These events are then leveraged to predict bi-directional optical flow. Guided by the estimated flow, we reconstruct intermediate frames. Next, we introduce EvFocusNet, which fuses the reconstructed frames while transforming the varying depth cues embedded in the CEFS, ultimately generating high-quality refocused results.

EvInterNet. To address the noise and spatial-temporal variant threshold issue, we adopt a data-driven approach and propose the CEFS-guided intermediate frame generation network, EvInterNet, which comprises two modules: Motion Filter and EvFlow. The Motion Filter is responsible for extracting motion-triggered events from the CEFS, while the EvFlow module leverages these features to estimate bi-directional optical flow and subsequently warp the boundary frames to synthesize the target intermediate frame $\mathbf{I}^\tau$.

[2]An illustrative example is provided in the supplementary video.

Since the difference between the two input frames, $\mathbf{I}^{\texttt{diff}}$, serves as an important cue for separating motion-triggered events, shown in Equation (8), the core challenge lies in effectively utilizing this cue to guide the event extraction process. This is particularly difficult due to the domain gap between image and event modalities. To tackle this issue, we introduce the Motion Filter, which is designed to separate motion-triggered events from other noise or sweep-triggered events. Given the proven effectiveness of attention mechanisms in multi-modal fusion [58], we employ multi-layer cross-attention blocks [64] to guide the filtering of motion-triggered events using image cues. The cross-attention block is defined as:

$$\mathcal{F}_e = \texttt{CrossAtten}(\mathcal{F}_{\texttt{diff}},\ \mathcal{F}_{\mathcal{E}}), \tag{9}$$

where $\mathcal{F}_{\texttt{diff}}$ and $\mathcal{F}_{\mathcal{E}}$ are the feature maps of the frame difference $\mathbf{I}^{\texttt{diff}}$ and CEFS $\mathcal{E}$, respectively. In the cross attention, the *key* and *value* vectors are derived from $\mathcal{F}_{\mathcal{E}}$, while the *query* vector from $\mathcal{F}_{\texttt{diff}}$.

After obtaining features corresponding to motion-triggered events, we leverage them to synthesize the desired intermediate frame. To ensure temporal consistency, we first convert the event features into optical flow, which is then used to warp the input frames. Inspired by the architecture of RIFE [19], the EvFlow module generates bi-directional optical flow from the extracted motion-triggered event features in two stages: *flow initialization* and *flow refinement*.

During flow initialization, we employ an EFBlock to produce a coarse flow estimate from the motion-triggered event features. To guide the synthesis towards the desired intermediate timestamp τ, a time embedding is introduced before the refinement stage, enabling the prediction of more accurate bi-directional flow. The refinement stage further enhances the flow fields using two additional EFBlocks. The overall process is expressed as:

$$f_{\tau\to0},\ f_{\tau\to1},\ M = \texttt{Flow}(\mathcal{F}_e,\ \mathbf{I}^1,\ \mathbf{I}^0,\ \tau), \tag{10}$$

where $f_{\tau\to0}$ and $f_{\tau\to1}$ denote the bi-directional optical flow, M is the fusion map, and $\texttt{Flow}$ represents the implicit function of the EvFlow module. The EFBlock follows the structure of IFBlock [19], but replaces convolution layers with attention blocks to better fuse event and image features across modalities. Once the optical flow and fusion map are estimated, a backward warping operation [19] is applied to generate the intermediate frame $\mathbf{I}^\tau$.

EvFocusNet. Since the generated intermediate frame $\mathbf{I}^\tau$ remains focused at the same depth as $\mathbf{I}^0$ and $\mathbf{I}^1$, we further refocus $\mathbf{I}^\tau$ to an arbitrary focal distance, guided by sweep-triggered events from the CEFS. Due to object motion, there are spatial discrepancies between the position of an object when it is in focus and its position at time τ, making it necessary to transform the texture information to align with the

Algorithm 1 Focus tracking with CEFS

Data: Thresholds μ and θ
Input: CEFS $\mathcal{E}$, frames $\mathbf{I}^0$ and $\mathbf{I}^1$, and ROI position $\mathbf{x}$
Result: Refocused frames $\{\mathbf{I}^{t_r}_{d_r}\}^N$
for $\mathbf{I}^{\mathtt{roi}} \in \{\mathbf{I}^0, \mathbf{I}^1\}$ $\quad d_r \leftarrow$ BINARYSEARCH($\mathbf{I}^{\mathtt{roi}}$)
for $t_r \in (t_0, t_1)$
  Reconstruct $\mathbf{I}^{t_r}_{d_r}$ with $d_r \leftarrow d_0 + \frac{(d_1-d_0)\cdot(t_r-t_0)}{(t_1-t_0)}$
  if $\mathbb{D}(\mathbf{I}^{\mathtt{roi}}_{d_n}) - \mathbb{D}(\mathbf{I}^{\mathtt{roi}}_{d_r}) > \theta$ **then**
    Reconstruct $\mathbf{I}^{t_r}_{d_r}$ with $d_r \leftarrow$ BINARYSEARCH($\mathbf{I}^{t_r}$)
  end if
end for
Function BINARYSEARCH($\mathbf{I}$)
$N \leftarrow 0, F \leftarrow 1$
while $F - N > \mu$
  $d_n, d_f \leftarrow F - (F-N)/2, N + (F-N)/2$
  Reconstruct $\mathbf{I}_{d_n}, \mathbf{I}_{d_f}$ with Equation (6)
  if $\mathbb{D}(\mathbf{I}_{d_n}) > \mathbb{D}(\mathbf{I}_{d_f})$ **then** $F \leftarrow d_f$ **else** $N \leftarrow d_n$
  $d_r \leftarrow (F+N)/2$
end while
return $d_r \leftarrow (F+N)/2$

object location. As CEFS encodes both temporal and spatial information, the motion-triggered events can be exploited to align textures to the desired position at time τ.

Because objects can come into focus at any point within the event stream, we first need to identify when the target depth is in focus using global temporal information across the entire sweep cycle. We then refine the texture alignment using local information centered around moving objects. Inspired by the ability of transformer architectures to model long-range pixel interactions [64], we propose the CEFS-guided refocusing network, EvFocusNet, which consists of multiple transformer blocks [64]. EvFocusNet fuses the event stream and reference frames to generate high-quality refocused images:

$$\mathbf{I}^{\tau}_{d} = \mathtt{Refocus}(\mathbf{I}^1,\ \mathbf{I}^0,\ \mathbf{I}^{\tau},\ \mathcal{E},\ d), \tag{11}$$

where d denotes the target focal distance, and `Refocus` represents the implicit function of EvFocusNet. As illustrated in Figure 3, we embed the focal distance using a sinusoidal encoding and broadcast it into the latent space for conditioning. By integrating EvInterNet and EvFocusNet, our framework enables the generation of temporally consistent intermediate frames that are refocused at arbitrary focal distances in dynamic scenes.

D. Focus tracking

Benefiting from CEFS, which encodes both continuous depth and motion information of a dynamic scene, our method allows refocusing at arbitrary focal distances. This capability enables post-capture focus tracking by selecting appropriate refocused focal distances. Following traditional camera focus tracking methodologies, our approach requires the position of the region of interest (ROI), typically specified by the user, to produce a video consistently focused on the target region.

As outlined in Algorithm 1, for each ROI patch in $\mathbf{I}^0$ and $\mathbf{I}^1$, we use binary search to determine the focal distance of ROI patch in the two frames, using the variance $\mathbb{D}$ of the reconstructed image intensity values to assess image sharpness. After obtaining the focal distance range from the start and end frames, we compute a linearly interpolated focal distance based on this initial range. Each frame is then reconstructed using Equation (6), producing a sequence of refocused frames for focus tracking. If any reconstructed frames are out of focus (*i.e.*, sharpness falls below that of the initial refocused frame by a threshold θ), we repeat the binary search to update the refocused focal distance.

E. Dynamic refocusing dataset

Existing refocusing datasets lack large-scale focal stack videos with corresponding event data. Lou *et al.* [33] utilize static focal stacks generated through rendering techniques. To address this limitation, we extend their data synthesis pipeline by incorporating dynamic elements, thereby introducing a dynamic refocusing dataset designed to leverage the rich temporal and spatial information embedded in event data for high-fidelity video refocusing.

We construct our dataset in two distinct ways, resulting in two subsets: `DF-Motion` and `DF-Shake`. The `DF-Motion` subset focuses primarily on object motion, combining diverse scenes and various object motion patterns. However, these synthetic scenes often lack realism compared to real-world scenarios. To ensure our method's applicability to real-world contexts, we additionally utilize light field data to synthesize detailed textures and realistic scenes for the testing `DF-Shake` subset, as light field data, capable of capturing comprehensive light ray information within a scene, enables the generation of highly realistic focal stacks [31]. The data synthesis methods are detailed as follows:

- `DF-Motion` comprises dynamic scenes rendered using Blender [6], emphasizing object motion. Each scene includes multiple distinct objects exhibiting diverse trajectories and motion patterns, both within horizontal planes and along the depth axis. For each scene, a focal stack video is generated using a shallow depth-of-field (DoF) camera setup (aperture f/1.2, focal length 100mm) that sweeps through the scene depth within a predefined duration.
- `DF-Shake` is synthesized from the Stanford Multiview Light Field (LiFF) Dataset [9]. Initially, a static focal stack is generated from light field data. Subsequently, for each frame within this focal stack, a randomly generated motion trajectory is introduced to simulate camera shake, transforming the static stack into a focal stack video.

After obtaining the focal stack videos, we input them into the event simulator DVS-voltmeter [32] to produce event streams. To enhance the model's generalization to diverse event sources, we randomly apply one of the six camera parameters (k_1 to k_6) available in the DVS-Voltmeter. Each parameter is sampled from the range [`Min`, `Max`], where `Min` $= [5.265, 20, 1e^{-4}, 1e^{-7}, 5e^{-9}, 1e^{-6}]$, based on calibrations from the DVS346 event camera models provided as preset configurations in DVS-Voltmeter [32], and `Max` is set to 5 times than `Min` values.

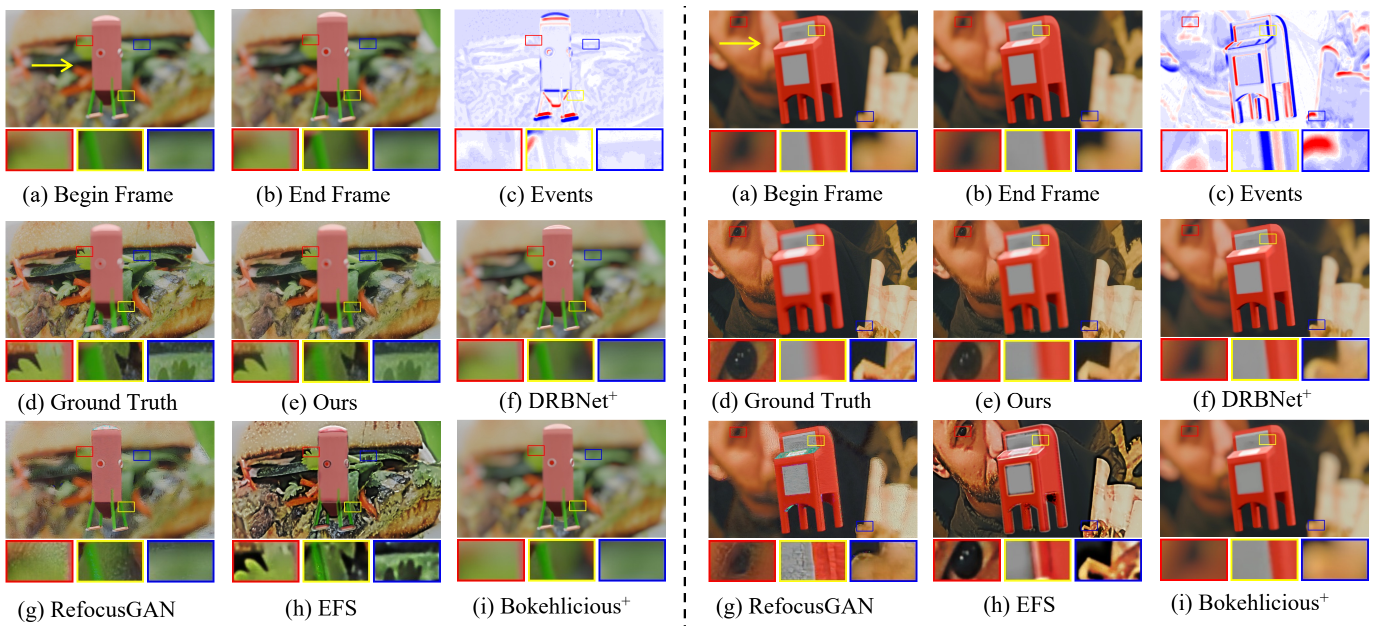


Fig. 4. Visual quality comparison with refocusing methods on the `DF-Motion` dataset. (a) Begin frame. (b) End frame. (c) Events. (d) Ground truth. (e)∼(i) Refocusing results of ours, DRBNet+ [44], RefocusGAN [46], EFS [33], and Bokehlicious+ [47]. Yellow arrows mark the target focal points. Close-up views are provided below each image. Additional results are available in the supplementary material.

Overall, our synthetic dataset consists of multiple image-event triplets: two adjacent frames with the same focal distance, and event streams triggered between these two frames. `DF-Motion` consists of 40 dynamic scenes, each containing 800 frames, with 30 scenes designated for training and 10 scenes reserved for testing. `DF-Shake` includes 30 dynamic scenes used exclusively for testing, each containing 50 frames. For each scene, we select three focal distances to conduct refocusing tests.[3]

F. Implementation details

We adopt a two-stage training strategy for EvInterNet and EvFocusNet. In the first stage, EvInterNet and EvFocusNet are trained separately by optimizing the following loss function:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{perc}}(\mathbf{I}^o, \mathbf{I}^{\text{gt}}) + \beta \cdot \mathcal{L}_1(\mathbf{I}^o, \mathbf{I}^{\text{gt}}), \tag{12}$$

where $\mathcal{L}_{\text{perc}}$ denotes the perceptual loss computed from a VGG-19 network [50] pretrained on ImageNet [45], and $\mathcal{L}_1$ denotes the mean absolute error. $\mathbf{I}^o$ represents the predicted output image, and $\mathbf{I}^{\text{gt}}$ denotes the corresponding ground truth. In our experiments, we set $\alpha = 3.5$ and $\beta = 70$.

EvInterNet and EvFocusNet are first trained jointly for 100 epochs using a learning rate of 5×10^{-5} and a batch size of 4. In the second stage, EvFocusNet is further trained for an additional 100 epochs while keeping EvInterNet frozen, using the same learning rate and batch size. Finally, we fine-tune both networks jointly with a learning rate of 5×10^{-5} and a batch size of 2 for another 60 epochs.

Our model is implemented in PyTorch and trained on an NVIDIA GeForce RTX 3090 GPU. Adam [23] is used as the optimizer throughout all training stages. During EvInterNet training, the ground-truth frame is selected at a random timestamp between the start and end of exposure to serve as the target temporal reconstruction point. For EvFocusNet, we randomly sample a focal distance from near zero to infinity as the supervision target. Events are pre-processed into event stacks following [60]. To improve robustness, random cropping is applied as the data augmentation strategy.

[3]More details about the data generation pipeline and setup can be found in the supplementary material.

TABLE II
QUANTITATIVE COMPARISONS ON THE `DF-MOTION` AND `DF-SHAKE` DATASET. ↑ (↓) INDICATES THE HIGHER (LOWER), THE BETTER THROUGHOUT THIS PAPER. THE BEST PERFORMANCES ARE HIGHLIGHTED IN **BOLD**.

	PSNR ↑	SSIM ↑	LPIPS ↓
`DF-Motion`			
DRBNet+ [44]	21.19	0.8332	0.3804
RefocusGAN [46]	22.45	0.6665	0.5402
Bokehlicious+ [47]	21.56	0.8255	0.3880
EFS [33]	27.06	0.8508	0.2969
Ours	**29.69**	**0.9316**	**0.1128**
`DF-Shake`			
DRBNet+ [44]	21.23	0.7757	0.4221
RefocusGAN [46]	19.87	0.6220	0.5742
Bokehlicious+ [47]	22.75	0.7892	0.4336
EFS [33]	25.77	0.8191	0.3218
Ours	**27.63**	**0.8609**	**0.2395**

IV. EXPERIMENTAL RESULTS

A. Quantitative comparison using synthetic data

We compare our method with three recent image-based refocusing approaches and one event-based method: DRBNet [44], RefocusGAN [46], Bokehlicious [47], and EFS [33].

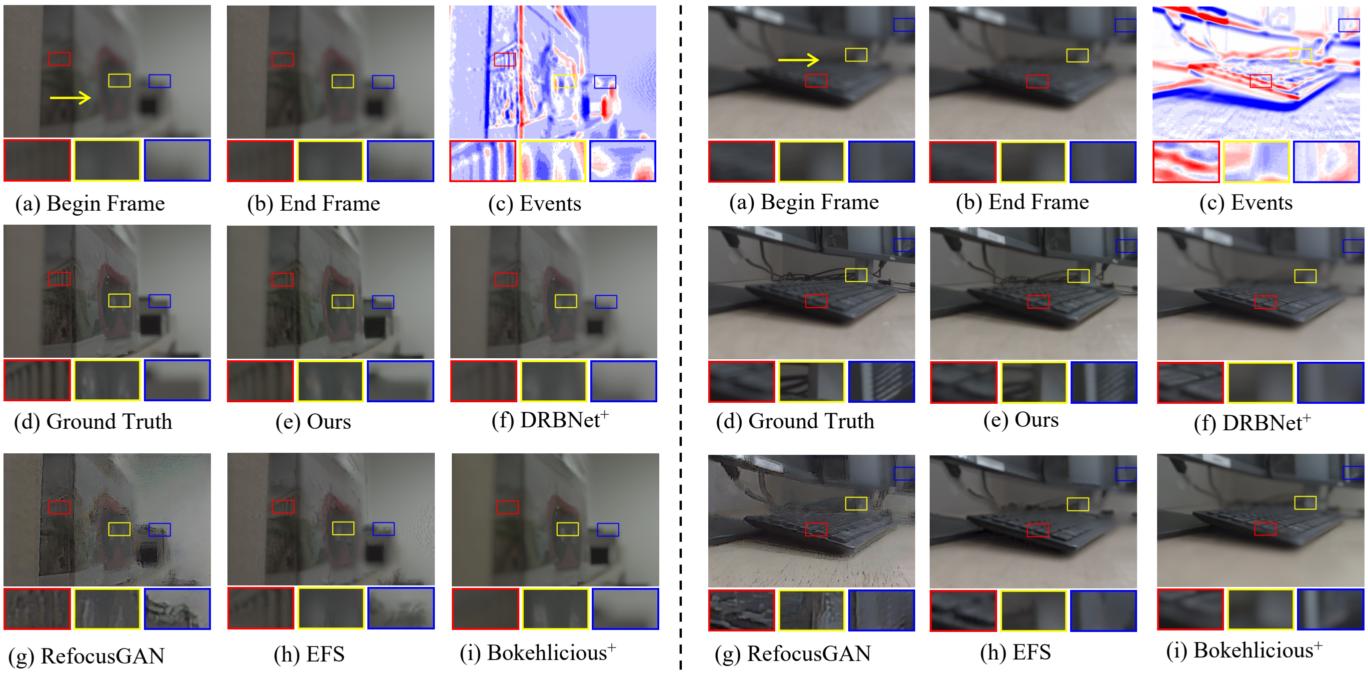


Fig. 5. Visual quality comparison with refocusing methods on the `DF-Shake` dataset. (a) Begin frame. (b) End frame. (c) Events. (d) Ground truth. (e)∼(i) Refocusing results of ours, DRBNet$^+$ [44], RefocusGAN [46], EFS [33], and Bokehlicious$^+$ [47]. Yellow arrows mark the target focal points. Close-up views are provided below each image. More results are available in the supplementary material.

Notably, since DRBNet [44] and Bokehlicious [47] are originally designed for defocus deblurring, we integrate them with BokehMe [38] for refocusing, and refer to the combined version as DRBNet$^+$ and Bokehlicious$^+$.

For a fair comparison, we fine-tuned all comparison methods, including EFS [33] and DRBNet [44], using our dataset. For training RefocusGAN, we used four distinct refocus parameter values: $\delta \in \{-9, -5, +5, +9\}$. The model was trained from scratch using 2250 focal stacks, each covering one-fifth of the images from our training set's focal stacks, sampled without overlap. All other settings were kept consistent with those described in the original manuscript [46]. Since EFS cannot directly accept depth as input to control the refocusing focal distance, we generate refocused images using its golden search method and then select the best-matched images.

We utilize Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) to measure the quality of restored images. Quantitative comparisons are presented in Table II, while qualitative comparisons are shown in Figure 4 and Figure 5. Our method consistently outperforms existing state-of-the-art methods across all evaluation metrics.

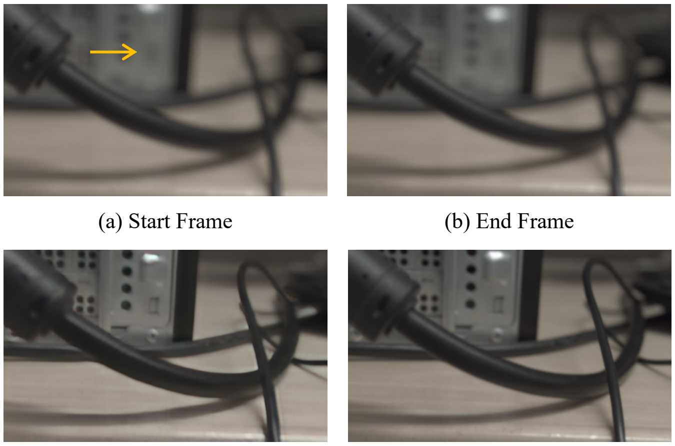
From the qualitative results, we observe that DRBNet$^+$ and Bokehlicious$^+$ struggle to recover fine texture details, and RefocusGAN tends to produce refocused outputs with reduced fidelity. In contrast, our method successfully restores high-frequency details encoded within the CEFS. Compared to the event-based method EFS [33], which suffers from artifacts and misalignment caused by entangled motion-triggered events, our approach mitigates these issues by explicitly separating motion-triggered and sweep-triggered events.

Furthermore, since our method is capable of generating intermediate frames, we conduct two-times and four-times frame interpolation experiments. As none of the comparative methods support intermediate frame synthesis directly, we combine each of them with the state-of-the-art frame interpolation method RIFE [19] for a fair comparison. Quantitative results on the `DF-Motion` dataset are reported in Table III, showing that our method achieves superior temporal consistency for refocused frames in dynamic scenes.

TABLE III
QUANTITATIVE COMPARISONS ON THE `DF-Motion` DATASET FOR INTERMEDIATE FRAME GENERATION. ALL COMPARED METHODS ARE INTERPOLATED BY RIFE [19].

	PSNR ↑	SSIM ↑	LPIPS ↓
2× interpolation			
DRBNet$^+$ [44]	21.09	0.8293	0.3895
RefocusGAN [46]	22.82	0.7000	0.5149
Bokehlicious$^+$ [47]	21.43	0.8254	0.3901
EFS [33]	26.45	0.8473	0.2992
Ours	**29.59**	**0.9410**	**0.1196**
4× interpolation			
DRBNet$^+$ [44]	21.07	0.8292	0.3905
RefocusGAN [46]	22.97	0.7129	0.5055
Bokehlicious$^+$ [47]	21.41	0.8256	0.3903
EFS [33]	26.34	0.8475	0.2997
Ours	**29.75**	**0.9465**	**0.1170**

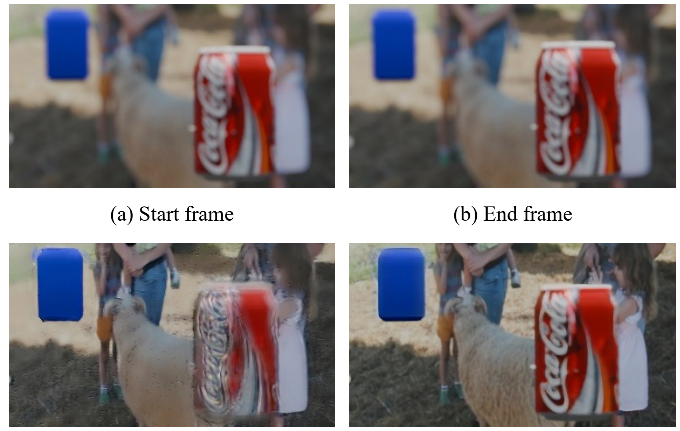
Motion analysis. As shown in Equation (8), the aggregated event signals approximate the summation of motion-triggered events. Importantly, our method does not rely on any specific type of motion—it is designed to handle a wide range of motion patterns, including both object movement and cam-



(a) Start Frame (b) End Frame

(c) Ours (d) GT

Fig. 6. An example illustrating a motion pattern unseen during training—camera moving outward from the scene on `DF-Shake` datatset. Given the start frame (a) and end frame (b), our method (c) successfully refocuses the image to the target focal distance, as indicated by the yellow arrow.



(a) Start frame (b) End frame

(c) TimeLens (d) Ours

Fig. 7. An example illustrating interpolation performance on the `DF-Motion` dataset. Given the start frame (a) and end frame (b), we compare TimeLens [57] (c) with our EvIntNet (d). Compared to TimeLens, our method successfully interpolates the image at the target timestamp without introducing artifacts.

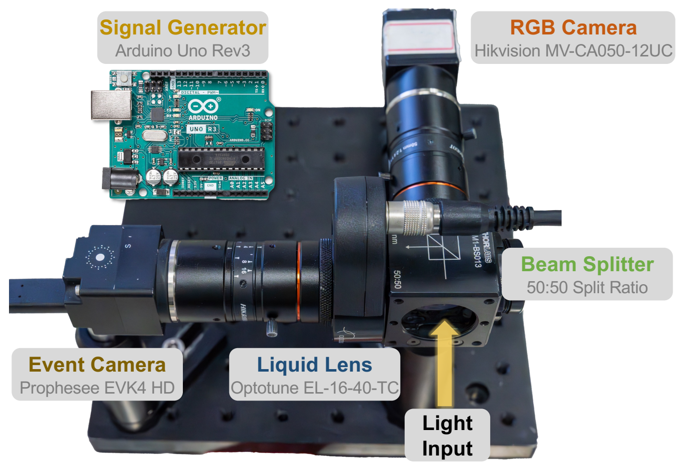


Fig. 8. The equipment setup we used to collect the real-world data.

era shake. To evaluate the generalization capability of our approach, we conduct experiments on motion patterns unseen during training. As illustrated in Figure 6, we test a scenario where the camera moves outward from the scene—a motion type not included in the training data. The results show that our method can still accurately refocus the image to the target focal distance, demonstrating strong robustness to diverse and unseen motion types.

Comparison with Timelens. To further demonstrate the effectiveness of our EvIntNet, we compared it with an existing event-based interpolation method, TimeLens [57], which is specifically designed to process motion-triggered events only. Unlike TimeLens [57], our EvIntNet explicitly separates the two types of events and utilizes only motion-triggered events for interpolation. In our experiment, we replaced EvIntNet with TimeLens [57] for the interpolation task. As shown in Figure 7, TimeLens [57] fails to accurately reconstruct intermediate frames. Quantitatively, TimeLens [57] achieves a PSNR of 24.52, whereas our method attains 29.75 on the `DF-Motion` dataset for $4\times$ interpolation.

B. Qualitative comparison using real data

To evaluate the effectiveness of our method in real-world scenarios, we construct a hybrid imaging platform consisting of a machine-vision camera (HIKVISION MV-CA050-12UC) and an event camera (PROPHESEE GEN4.0), as shown in Figure 8. As illustrated in Figure 1(a), a liquid lens (Optotune EL-16-40-TC) is attached to the event camera to enable continuous focal sweep. In the previous work [33], the focal sweep was performed manually with varying speeds, yet the high temporal resolution of event cameras ensured negligible impact on event generation and reconstruction quality; the same conclusion holds here, as the liquid-lens modulation frequency used in our system (up to 100 Hz) remains far below the intrinsic temporal resolution of the event sensor. Unlike conventional focusing mechanisms that translate lens groups along the optical axis and introduce optical "breathing" (*i.e.*, variations in magnification or field of view during refocusing), a liquid lens modulates focus by deforming a fluid–membrane interface, thereby largely suppressing geometric scaling effects [10], [59]. Both cameras are equipped with 50 mm F/1.8 lenses and are co-aligned using a beam splitter to maintain a shared optical path. An external signal generator (Arduino Uno Rev3) provides temporal synchronization, and standard camera calibration procedures are applied to ensure spatial alignment between modalities.

Visual comparisons on real-world data are shown in Figure 9, including indoor scenes (first row), outdoor environments (second row), and cases involving non-rigid motion (third row). Across all three categories, our method reconstructs refocused images with fine textures and sharp details, even where the input frames exhibit strong defocus. By contrast, image-based approaches fail to recover high-frequency structure and produce blurred or over-smoothed outputs, while EFS [33], though more tolerant to defocus, introduces artifacts due to motion-induced events that are not explicitly separated.

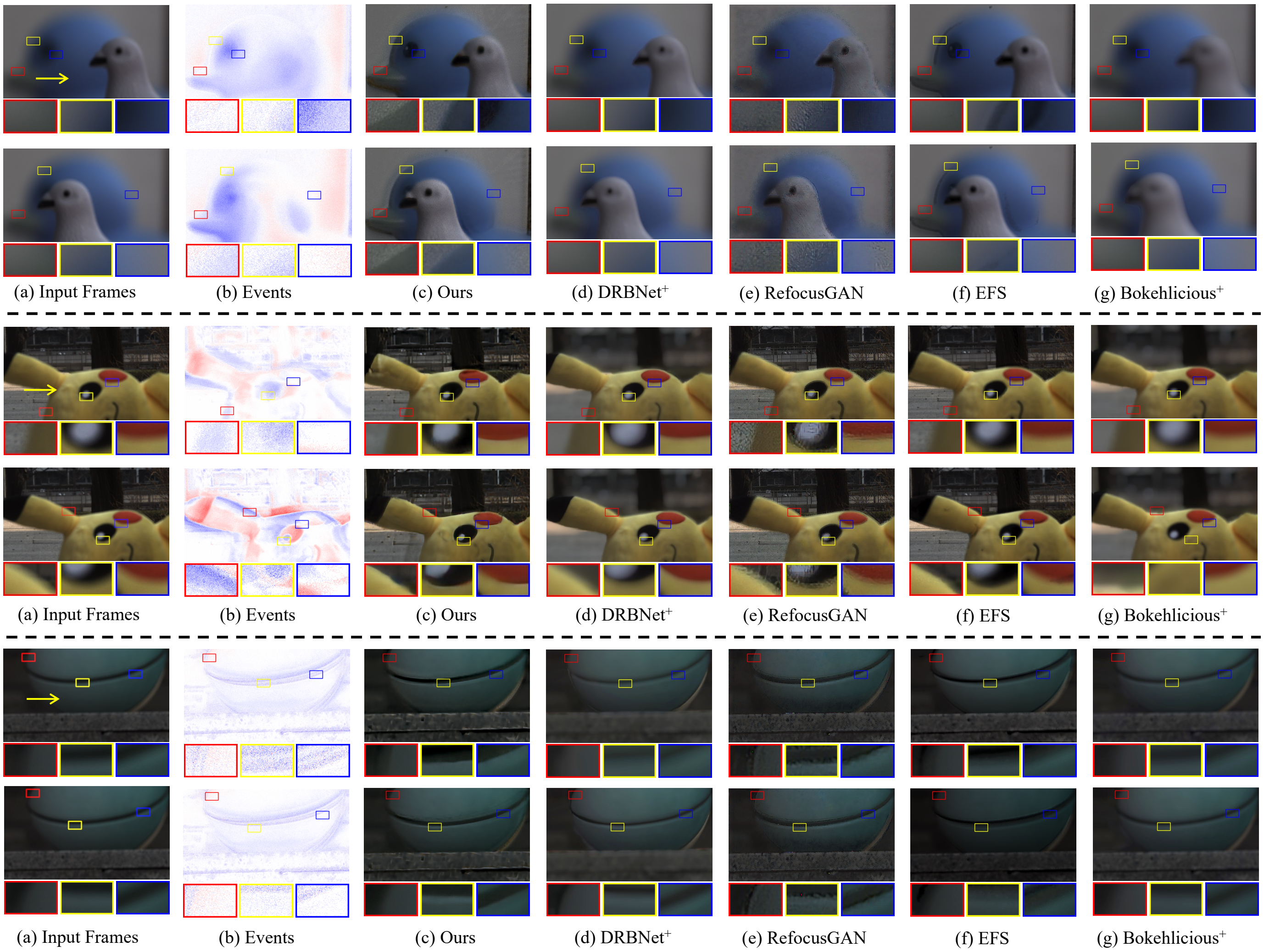


Fig. 9. Visual comparison on the real-captured data. (a) Input frames. (b) Events. (c)∼(g) Refocusing results generated by ours, DRBNet$^+$ [44], RefocusGAN [46], EFS [33], and Bokehlicious$^+$ [47]. For each example, the top row displays the begin frame and its corresponding refocused results, while the bottom row shows the end frame and corresponding refocused results. Yellow arrows mark the target focal points. Close-up views are provided below each image to highlight refocusing details. Additional results are provided in the supplementary material.

Leveraging disentangled sweep-triggered and motion-triggered events, our approach consistently delivers clearer refocusing and generalizes effectively across a range of real-world scene complexities.

C. Application

Focus tracking. Our approach supports focus tracking based on the position of a given region of interest (ROI). An example is shown in Figure 10: starting from frames focused at a near-zero focal distance (*i.e.*, defocused frames), we refocus the sequence to continuously track the helmet as it moves from a far to a near position. The results demonstrate that our method can effectively perform focus tracking on moving objects across varying focal distances. Compared to other methods, our approach maintains better temporal consistency and introduces fewer artifacts. Video results showcasing focus tracking in dynamic scenes are provided in the supplementary material.

TABLE IV
QUANTITATIVE COMPARISONS FOR FOCUS TRACKING ON THE DF-MOTION DATASET AND ALL-IN-FOCUS IMAGING ON THE DF-SHAKE DATASET. IN THE FOCUS TRACKING TASK, ALL COMPARED METHODS ARE INTERPOLATED BY RIFE [19].

	PSNR ↑	SSIM ↑	LPIPS ↓
Focus tracking			
DRBNet$^+$ [44]	24.79	0.9522	0.1059
RefocusGAN [46]	25.61	0.9596	0.0937
EFS [33]	29.29	0.9719	0.0470
Ours	**30.25**	**0.9826**	**0.0380**
All-in-focus imaging			
DRBNet [44]	20.53	0.6648	0.3162
RefocusGAN [46]	21.31	0.5333	0.6421
EFS [33]	20.00	0.6267	0.3950
Ours	**23.58**	**0.7367**	**0.2934**

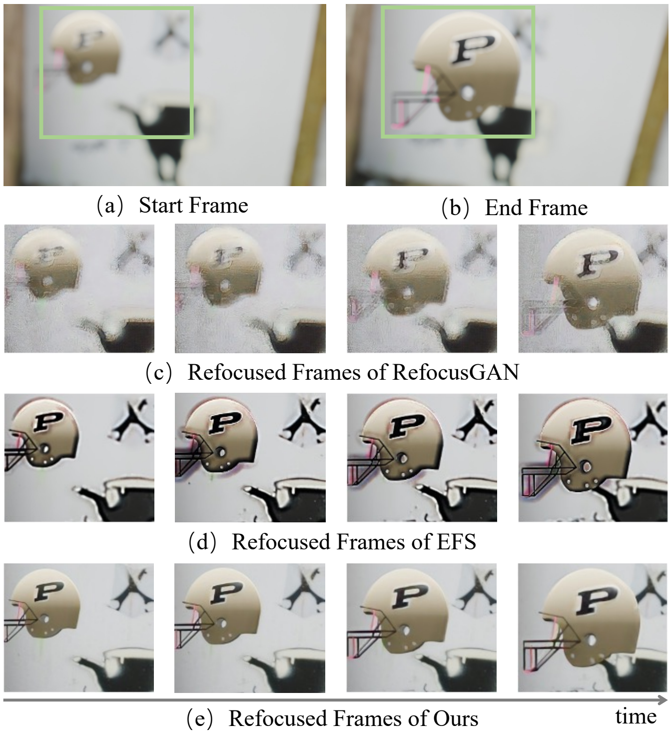


Fig. 10. An example of focus tracking on `DF-Motion` dataset. (a) Start frame. (b) End frame. (c)∼(e) The cropped generated intermediate frames of RefocusGAN [46], EFS [33], and ours, which focus on the helmet, with the focal distance shifting from far to near.

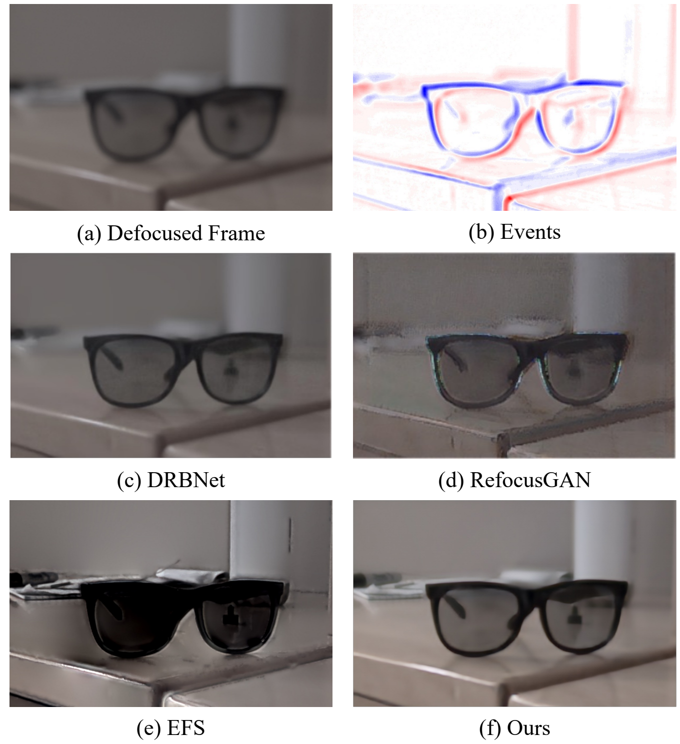


Fig. 11. Visual quality comparison of all-in-focus imaging on `DF-Shake` dataset. (a) Defocused image. (b) Events. (c)∼(f) All-in-focus images recovered by DRBNet [44], RefocusGAN [46], EFS [33] and our method.

To evaluate whether the predicted image is properly focused in regions that are expected to be in focus, we adopt a reference-based local sharpness assessment. Specifically, we manually define a region of interest (ROI) Ω, and generate a corresponding binary mask as follows:

$$\mathcal{M}(x,y) = \begin{cases} 1, & \text{if } (x,y) \in \Omega \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

Here, $\mathcal{M}$ indicates the high-frequency regions that are expected to be accurately refocused. We then compute focus-aware quality metrics, exclusively within the masked region. For each scene in the `DF-Motion` dataset, we select two objects as the targets for focus tracking. The quantitative results are reported in Table IV, showing that our method achieves the highest performance in terms of localized focus accuracy, consistently outperforming other approaches.

All-in-focus imaging. Since our method can generate an image focal stack at any specific timestamp, it naturally extends to support all-in-focus imaging as a downstream task. Quantitative results are presented in Table IV, and qualitative comparisons are shown in Figure 11. Across all metrics, our method achieves the best performance.

As illustrated in Figure 11, our approach leverages CEFS to guide the warping of texture information toward the target timestamp, producing high-quality all-in-focus images. In contrast, EFS [33] struggles to handle motion during the focal sweep, resulting in noticeable artifacts. Among image-based methods, DRBNet [44] and RefocusGAN [46] fail to recover sharp details from defocused frames. It is worth noting that while DRBNet is capable of directly generating all-in-focus images, we only utilize the defocus removal module of RefocusGAN for a fair comparison.

D. Ablation studies

To evaluate the effectiveness of each component in our framework, we conduct four ablation studies. First, we remove the Motion Filter to assess its contribution in extracting motion-triggered events (denoted as “W/o Motion Filter”). Second, we exclude the EvFlow module to evaluate the importance of intermediate frame generation, replacing it with direct embedding of focal distance and time into EvFocusNet (denoted as “W/o EvFlow”). Third, to examine the impact of the training scheme, we reverse the order of EvInterNet and EvFocusNet (denoted as “Reverse”). Finally, we investigate the necessity of staged training by jointly training both modules from scratch (denoted as “End2End”).

As shown in Table V, our complete model achieves the best performance in both refocusing and interpolation tasks. The detailed analysis of each ablation study is listed as follows:

- W/o Motion Filter: Although the Motion Filter is a lightweight four-layer cross-attention block, removing it causes a clear degradation under $2\times$ interpolation, confirming that performance drops mainly due to the loss of motion-filtering capability rather than parameter reduction. The resulting inferior intermediate frame restoration further propagates error, ultimately lowering overall interpolation quality.

TABLE V
QUANTITATIVE RESULTS OF ABLATION STUDY.

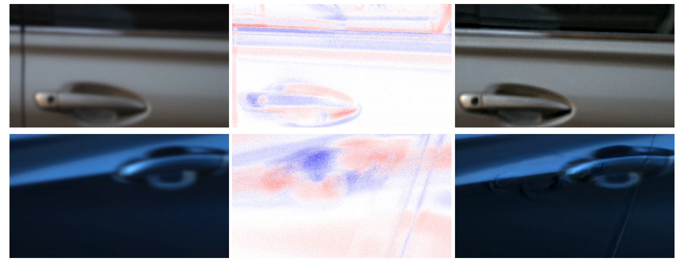
	PSNR ↑	SSIM ↑	LPIPS ↓
DF-Motion			
W/o Motion Filter	29.65	0.9311	0.1134
W/o EvFlow	23.25	0.8723	0.1718
Reverse	23.70	0.5514	0.6550
End2End	23.85	0.8731	0.2040
Our complete model	**29.69**	**0.9316**	**0.1128**
2× interpolation on DF-Motion			
W/o Motion Filter	26.97	0.9163	0.1394
W/o EvFlow	25.54	0.9062	0.1524
Reverse	23.62	0.4949	0.6822
End2End	22.00	0.8293	0.2231
Our complete model	**29.59**	**0.9410**	**0.1196**

- W/o EvFlow: Without the EvFlow module, EvFocusNet is required to adjust both the focal distance and time simultaneously. This additional complexity hinders the performance of EvFocusNet, limiting its effectiveness in achieving optimal results.
- Reverse: Generating intermediate frames is comparatively easier than adjusting the focal distance. Therefore, reversing the order of EvInterNet and EvFocusNet leads to a decline in overall performance.
- End2End: When learning intermediate frame generation and focal distance adjustment simultaneously, both EvInterNet and EvFocusNet struggle to focus effectively on reconstruction. Training them separately in the first stage allows each network to develop an initial capability for generating frames. Consequently, directly training the two networks jointly degrades their individual performance, thereby adversely affecting the overall performance of the system.

V. CONCLUSION

This paper introduces the Coded Event Focal Stack (CEFS), a novel technique for precise refocusing in dynamic scenes by capturing event streams during a modulated focal sweep. We propose an event-guided dynamic refocusing framework that effectively disentangles motion and depth information by leveraging the high temporal resolution of event cameras. Guided by CEFS, our method generates continuous intermediate frames refocused at arbitrary focal distances. To support real-world evaluation, we construct a hybrid camera system equipped with a liquid lens for programmable focal sweeping. Experimental results on real-captured data demonstrate that our method consistently outperforms existing state-of-the-art approaches in both visual quality and temporal consistency.

Limitation. While CEFS performs well in scenes with moderate motion, its performance degrades in the presence of noticeable motion blur in the RGB frame, which leads to less reliable event decomposition and visible artifacts in the refocused output. As illustrated in Figure 12, clean refocusing is achieved in the case without motion blur, whereas the presence of blur produces observable reconstruction artifacts.



(a) Defocused Frame (b) Events (c) Refocused Frame

Fig. 12. Examples of high-speed motion. (a) Defocused frame. The first row contains frames without motion blur, while the second row contains frames with noticeable motion blur. (b) Corresponding event data. (c) Refocused results. When motion blur is absent, our method effectively restores fine structural details. However, when the input frames exhibit motion blur, artifacts appear in the refocused output. Addressing motion-blur-induced degradation remains an open direction for future work.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (Grant No. 62136001, 62401021), Beijing Natural Science Foundation (Grant No. L233024), and Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012).

REFERENCES

[1] Abuolaim, A., Brown, M.S.: Defocus deblurring using dual-pixel data. In: Proc. of European Conference on Computer Vision. pp. 111–126 (2020)
[2] Abuolaim, A., Delbracio, M., Kelly, D., Brown, M.S., Milanfar, P.: Learning to reduce defocus blur by realistically modeling dual-pixel data. In: Proc. of International Conference on Computer Vision. pp. 2289–2298 (2021)
[3] Alzayer, H., Abuolaim, A., Chan, L.C., Yang, Y., Lou, Y.C., Huang, J.B., Kar, A.: DC^2: Dual-camera defocus control by learning to refocus. In: Proc. of Computer Vision and Pattern Recognition. pp. 21488–21497 (2023)
[4] Baltag, O.: History of automatic focusing reflected by patents. Science **3**(1), 1–17 (2015)
[5] Bao, Y., Sun, L., Ma, Y., Gu, D., Wang, K.: Improving fast auto-focus with event polarity. Optics Express **31**(15), 24025–24044 (2023)
[6] Blender Foundation: The Blender project - free and open 3D creation software, https://www.blender.org, accessed: 2022-11-04
[7] Chen, H., Teng, M., Shi, B., Wang, Y., Huang, T.: A residual learning approach to deblur and generate high frame rate video with an event camera. IEEE Transactions on Multimedia **25**, 5826–5839 (2023)
[8] Chen, S., Guo, M.: Live demonstration: CeleX-V: A 1m pixel multi-mode event-based sensor. In: Proc. of Computer Vision and Pattern Recognition Workshops. pp. 1682–1683 (2019)
[9] Dansereau, D.G., Girod, B., Wetzstein, G.: LiFF: Light field features in scale and depth. In: Proc. of Computer Vision and Pattern Recognition. pp. 8042–8051 (2019)
[10] Dorronsoro, C., Barcala, X., Gambra, E., Akondi, V., Sawides, L., Marrakchi, Y., Rodriguez-Lopez, V., Benedi-Garcia, C., Vinas, M., Lage, E., et al.: Tunable lenses: dynamic characterization and fine-tuned control for high-speed applications. Optics express **27**(3), 2085–2100 (2019)
[11] Dowski, E.R., Cathey, W.T.: Extended depth of field through wave-front coding. Applied Optics **34**(11), 1859–1866 (1995)
[12] Duan, P., Wang, Z.W., Shi, B., Cossairt, O., Huang, T., Katsaggelos, A.K.: 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)
[13] D'Andrès, L., Salvador, J., Kochale, A., Süsstrunk, S.: Non-parametric blur map regression for depth of field extension. IEEE Transactions on Image Processing **25**(4), 1660–1673 (2016)

[14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)

[15] Guo, C., Ma, Z., Guo, X., Li, W., Qi, X., Zhao, Q.: Fast auto-focusing search algorithm for a high-speed and high-resolution camera based on the image histogram feature function. Applied Optics **57**(34), F44–F49 (2018)

[16] Hach, T., Steurer, J., Amruth, A., Pappenheim, A.: Cinematic bokeh rendering for real scenes. In: Proc. of European Conference on Visual Media Production. pp. 1–10 (2015)

[17] Horn, B.K.P.: Focusing. Tech. rep., MIT (1968)

[18] Hu, Y., Liu, S.C., Delbruck, T.: v2e: From video frames to realistic DVS events. In: Proc. of Computer Vision and Pattern Recognition Workshops. pp. 1312–1321 (2021)

[19] Huang, Z., Zhang, T., Heng, W., Shi, B., Zhou, S.: Real-time intermediate flow estimation for video frame interpolation. In: Proc. of European Conference on Computer Vision. pp. 624–642 (2022)

[20] Karaali, A., Jung, C.R.: Edge-based defocus blur estimation with adaptive scale selection. IEEE Transactions on Image Processing **27**(3), 1126–1137 (2017)

[21] Kehtarnavaz, N., Oh, H.J.: Development and real-time implementation of a rule-based auto-focus algorithm. Real-Time Imaging **9**(3), 197–203 (2003)

[22] Kim, H., Richardt, C., Theobalt, C.: Video depth-from-defocus. In: Proc. of International Conference on 3D Vision. pp. 370–379 (2016)

[23] Kingma, D.P., Ba, J.: ADAM: A method for stochastic optimization. In: Proc. of International Conference on Learning Representations (2015)

[24] Kuthirummal, S., Nagahara, H., Zhou, C., Nayar, S.K.: Flexible depth of field photography. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(1), 58–71 (2010)

[25] Lee, J., Lee, S., Cho, S., Lee, S.: Deep defocus map estimation using domain adaptation. In: Proc. of Computer Vision and Pattern Recognition. pp. 12222–12230 (2019)

[26] Lee, J., Son, H., Rim, J., Cho, S., Lee, S.: Iterative filter adaptive network for single image defocus deblurring. In: Proc. of Computer Vision and Pattern Recognition. pp. 2034–2042 (2021)

[27] Levin, A., Fergus, R., Durand, F., Freeman, W.T.: Image and depth from a conventional camera with a coded aperture. ACM Transactions on Graphics (Proc. of ACM SIGGRAPH) **26**(3), 70 (2007)

[28] Levin, A., Hasinoff, S.W., Green, P., Durand, F., Freeman, W.T.: 4D frequency analysis of computational cameras for depth of field extension. ACM Transactions on Graphics (Proc. of ACM SIGGRAPH) **28**(3), 1–14 (2009)

[29] Levin, A., Sand, P., Cho, T.S., Durand, F., Freeman, W.T.: Motion-invariant photography. ACM Transactions on Graphics (TOG) **27**(3), 1–9 (2008)

[30] Levoy, M., Ng, R., Adams, A., Footer, M., Horowitz, M.: Light field microscopy. ACM Transactions on Graphics (Proc. of ACM SIGGRAPH) **25**(3), 924–934 (2006)

[31] Lin, H., Chen, C., Kang, S.B., Yu, J.: Depth recovery from light field using focal stack symmetry. In: Proc. of International Conference on Computer Vision (2015)

[32] Lin, S., Ma, Y., Guo, Z., Wen, B.: DVS-Voltmeter: Stochastic process-based event simulator for dynamic vision sensors. In: Proc. of European Conference on Computer Vision. pp. 578–593 (2022)

[33] Lou, H., Teng, M., Yang, Y., Shi, B.: All-in-focus imaging from event focal stack. In: Proc. of Computer Vision and Pattern Recognition. pp. 17366–17375 (2023)

[34] Miau, D., Cossairt, O., Nayar, S.K.: Focal sweep videography with deformable optics. In: Proc. of International Conference on Computational Photography. pp. 1–8 (2013)

[35] Ng, R.: Digital light field photography. Stanford University (2006)

[36] Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y.: Bringing a blurry frame alive at high frame-rate with an event camera. In: Proc. of Computer Vision and Pattern Recognition. pp. 6820–6829 (2019)

[37] Park, J., Tai, Y.W., Cho, D., So Kweon, I.: A unified approach of multi-scale deep and hand-crafted features for defocus estimation. In: Proc. of Computer Vision and Pattern Recognition. pp. 1736–1745 (2017)

[38] Peng, J., Cao, Z., Luo, X., Lu, H., Xian, K., Zhang, J.: BokehMe: When neural rendering meets classical rendering. In: Proc. of Computer Vision and Pattern Recognition. pp. 16283–16292 (2022)

[39] Qu, Z., Zou, Z., Boominathan, V., Chakravarthula, P., Pediredla, A.: Event fields: Capturing light fields at high speed, resolution, and dynamic range. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26910–26920 (2025)

[40] Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(6), 1964–1980 (2019)

[41] Ruan, L., Bálint, M., Bemana, M., Wolski, K., Seidel, H.P., Myszkowski, K., Chen, B.: Self-supervised video defocus deblurring with atlas learning. In: Proc. of ACM SIGGRAPH (2024)

[42] Ruan, L., Bemana, M., Seidel, H.p., Myszkowski, K., Chen, B.: Revisiting image deblurring with an efficient convnet. arXiv preprint arXiv:2302.02234 (2023)

[43] Ruan, L., Chen, B., Li, J., Lam, M.L.: AIFNet: All-in-focus image restoration network using a light field-based dataset. IEEE Transactions on Computational Imaging **7**, 675–688 (2021)

[44] Ruan, L., Chen, B., Li, J., Lam, M.: Learning to deblur using light field generated and real defocus images. In: Proc. of Computer Vision and Pattern Recognition. pp. 16304–16313 (2022)

[45] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)

[46] Sakurikar, P., Mehta, I., Balasubramanian, V.N., Narayanan, P.: RefocusGAN: Scene refocusing using a single image. In: Proc. of European Conference on Computer Vision. pp. 497–512 (2018)

[47] Seizinger, T., Vasluianu, F.A., Conde, M., Wu, Z., Timofte, R.: Bokehlicious: Photorealistic bokeh rendering with controllable apertures. In: Proc. of International Conference on Computer Vision (2025)

[48] Serrano-Gotarredona, T., Linares-Barranco, B.: A 128$\times$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)

[49] Shi, J., Xu, L., Jia, J.: Just noticeable defocus blur detection and estimation. In: Proc. of Computer Vision and Pattern Recognition. pp. 657–665 (2015)

[50] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proc. of International Conference on Learning Representations (2015)

[51] Śliwiński, P., Wachel, P.: A simple model for on-sensor phase-detection autofocusing algorithm. Journal of Computer and Communications **1**(06), 11 (2013)

[52] Son, H., Lee, J., Cho, S., Lee, S.: Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In: Proc. of International Conference on Computer Vision. pp. 2642–2650 (2021)

[53] Teng, M., Lou, H., Yang, Y., Huang, T., Shi, B.: Hybrid all-in-focus imaging from neuromorphic focal stack. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 10124–10137 (2024)

[54] Teng, M., Zhou, C., Lou, H., Shi, B.: NEST: Neural event stack for event-based image enhancement. In: Proc. of European Conference on Computer Vision. pp. 660–676 (2022)

[55] Tulyakov, S., Bochicchio, A., Gehrig, D., Georgoulis, S., Li, Y., Scaramuzza, D.: Time Lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In: Proc. of Computer Vision and Pattern Recognition. pp. 17755–17764 (2022)

[56] Tulyakov, S., Gehrig, D., Georgoulis, S., Erbach, J., Gehrig, M., Li, Y., Scaramuzza, D.: Time Lens: Event-based video frame interpolation. In: Proc. of Computer Vision and Pattern Recognition. pp. 16155–16164 (2021)

[57] Tulyakov, S., Gehrig, D., Georgoulis, S., Erbach, J., Gehrig, M., Li, Y., Scaramuzza, D.: TimeLens: Event-based video frame interpolation. In: Proc. of Computer Vision and Pattern Recognition. pp. 16155–16164 (2021)

[58] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proc. of Conference and Workshop on Neural Information Processing Systems (2017)

[59] Wang, D., Li, Z.S., Zheng, Y., Zhao, Y.R., Liu, C., Xu, J.B., Zheng, Y.W., Huang, Q., Chang, C.L., Zhang, D.W., et al.: Liquid lens based holographic camera for real 3d scene hologram acquisition using end-to-end physical model-driven network. Light: Science & Applications **13**(1), 62 (2024)

[60] Wang, L., Ho, Y.S., Yoon, K.J., et al.: Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In: Proc. of Computer Vision and Pattern Recognition. pp. 10081–10090 (2019)

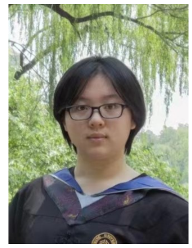
[61] Weng, W., Zhang, Y., Xiong, Z.: Event-based video reconstruction using transformer. In: Proc. of International Conference on Computer Vision. pp. 2563–2572 (2021)

[62] Yang, Z., Lian, W., Lai, S.: BokehOrNot: Transforming bokeh effect with image transformer and lens metadata embedding. In: Proc. of Computer Vision and Pattern Recognition. pp. 1542–1550 (2023)
[63] Yousefi, S., Rahman, M., Kehtarnavaz, N.: A new auto-focus sharpness function for digital and smart-phone cameras. IEEE Transactions on Consumer Electronics **57**(3), 1003–1009 (2011)
[64] Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proc. of Computer Vision and Pattern Recognition. pp. 5728–5739 (2022)
[65] Zhang, X., Yu, L., Yang, W., Liu, J., Xia, G.S.: Generalizing event-based motion deblurring in real-world scenarios. In: Proc. of Computer Vision and Pattern Recognition (2023)
[66] Zhao, W., Wei, F., He, Y., Lu, H.: United defocus blur detection and deblurring via adversarial promoting learning. In: Proc. of European Conference on Computer Vision. pp. 569–586 (2022)
[67] Zhou, C., Miau, D., K. Nayar, S.: Focal sweep camera for space-time refocusing. Tech. rep., Columbia University (2012)
[68] Zhou, C., Teng, M., Han, J., Xu, C., Shi, B.: Delieve-net: Deblurring low-light images with light streaks and local events. In: Proc. of Computer Vision and Pattern Recognition Workshops (2021)

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, image enhancement and video generation. He has served as a reviewer for CVPR, ICCV, ECCV, TPAMI, TIP, *etc*.

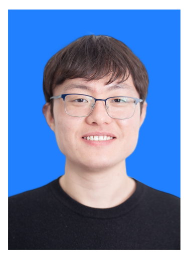
Suhang Xuan is currently an undergraduate student at the School of Electronics Engineering and Computer Science, Peking University, Beijing, China, expecting to receive his B.S. degree in 2026. His research interests are focused on computer vision, neuromorphic cameras, and model compression.

Zhiang Yan is currently an undergraduate student at the School of Electronics Engineering and Computer Science, Peking University, Beijing, China, expecting to receive his B.S. degree in 2026. His research interests are focused on computational photography, multi-modal learning and brain-inspired models.



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.

Boyu Li is currently a Ph.D. student in the School of Computer Science of Peking University. He received the B.S. degree from Peking University in 2023. His research interests include event-based vision and related topics.



Bin Fan received the B.S. degree, M.E. degree, and Ph.D. degree from Northwestern Polytechnical University (NPU), Xi'an, China, in 2016, 2019, and 2023, respectively. He is currently a Professor with the School of Electronics and Information at NPU. He was a Boya postdoctoral researcher at Peking University, Beijing, China. He was selected to the CVPR 2022 Doctoral Consortium and co-organized tutorials at CVPR 2023 and ACCV 2022. His research interests include computer vision, computational photography, and deep learning.

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.

Coded Event Focal Stack for Continuous Refocusing in Dynamic Scene Supplementary Material

Minggui Teng†, Suhang Xuan†, Zhiang Yan, Hanyue Lou, Boyu Li, Bin Fan, Boxin Shi‡, *Senior Member, IEEE*

I. Dataset

A. Training synthetic dataset

Inspired by the data generation approach in EFS [3], our pipeline for generating the training dataset is illustrated in Figure 1. To create the dataset, we first randomly select geometric objects, scale them, and distribute them at various depths across 30 distinct scenes. Next, we introduce dynamic elements to each scene by assigning random motion trajectories to the objects and camera, including translation, rotation, and shake. To better align the data distribution with real-world images, we apply textures sampled from the MS-COCO dataset [2] onto the surfaces of the objects. These textures are selected from the *2017 Val images*, a subset of MS-COCO. For each scene, we sample 50 depth-time paired points to render the high frame rate videos. The depth values are sampled from near to far, while the time steps are evenly sampled across the time interval. We select a subset of frames from the videos to serve as input frames.

After generating the high frame rate focal stack videos, we use DVS-Voltmeter [1] to produce event streams based on these rendered images. To enhance the model's generalization to diverse event sources, we randomly apply one of the six camera parameters (k_1 to k_6) available in the DVS-Voltmeter. Each parameter is sampled from the range [`Min`, `Max`], where `Min` $= [5.265, 20, 1e^{-4}, 1e^{-7}, 5e^{-9}, 1e^{-6}]$, based on calibrations from the DVS346 event camera models provided as preset configurations in DVS-Voltmeter [1], and `Max` is set to 5 times than `Min` values.

B. DF-Motion and DF-Shake dataset

Our `DF-Shake` and `DF-Motion` datasets feature objects with diverse trajectories and motion patterns, including movements along horizontal planes and the depth axis. Example scenes illustrating these characteristics are provided in the supplementary video.

† Contributed equally to this work as first authors.
‡ Corresponding author: shiboxin@pku.edu.cn
Minggui Teng, Suhang Xuan, Zhiang Yan, Hanyue Lou, 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.
Bin Fan is with School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China

TABLE I
THE QUANTITY OF MULTIPLY-ACCUMULATE OPERATIONS (MACS) AND THE COUNT OF NETWORK PARAMETERS (#PARAM) ACROSS ALL COMPARATIVE METHODS.

	MACs (G)	#Param (M)
DRBNet$^+$ [5]	820.04	12.12
RefocusGAN [6]	1610.87	22.76
EFS [3]	65947.75	4.90
Ours	3160.52	19.95

II. Experimental Details

For a fair comparison, we fine-tuned all comparison methods, including EFS [3] and DRBNet [5], using our dataset. As the code for RefocusGAN [6] is not publicly available, we implemented the model based on the instructions in the manuscript and trained it on our dataset.

DRBNet. DRBNet [5] is a defocus deblurring method that cannot perform image refocusing. To address this limitation, we integrated DRBNet [5] with BokehMe [4], a method capable of rendering all-in-focus images at specified depths. This combination allows the enhanced framework DRBNet$^+$ to perform image refocusing.

RefocusGAN. During training, we used four distinct refocus parameter values: $\delta \in \{-9, -5, +5, +9\}$. The model was trained from scratch using 2250 focal stacks, each covering one-fifth of the images from our training set's focal stacks, sampled without overlap. All other settings were kept consistent with those described in the original manuscript [6].

EFS. Since EFS cannot directly accept depth as input to control the refocusing focal distance, we generate refocused images using its golden search method and then select the best-matched images.

III. Network Parameters

We list multiply-accumulate operations and network parameters of comparative methods in Table I.

IV. CEFS Visualization

We provide visual examples illustrating the focal-sweep behavior in the supplementary video. The corresponding frames are extracted from the video and presented in Figure 2. Please see the video for better illustration.

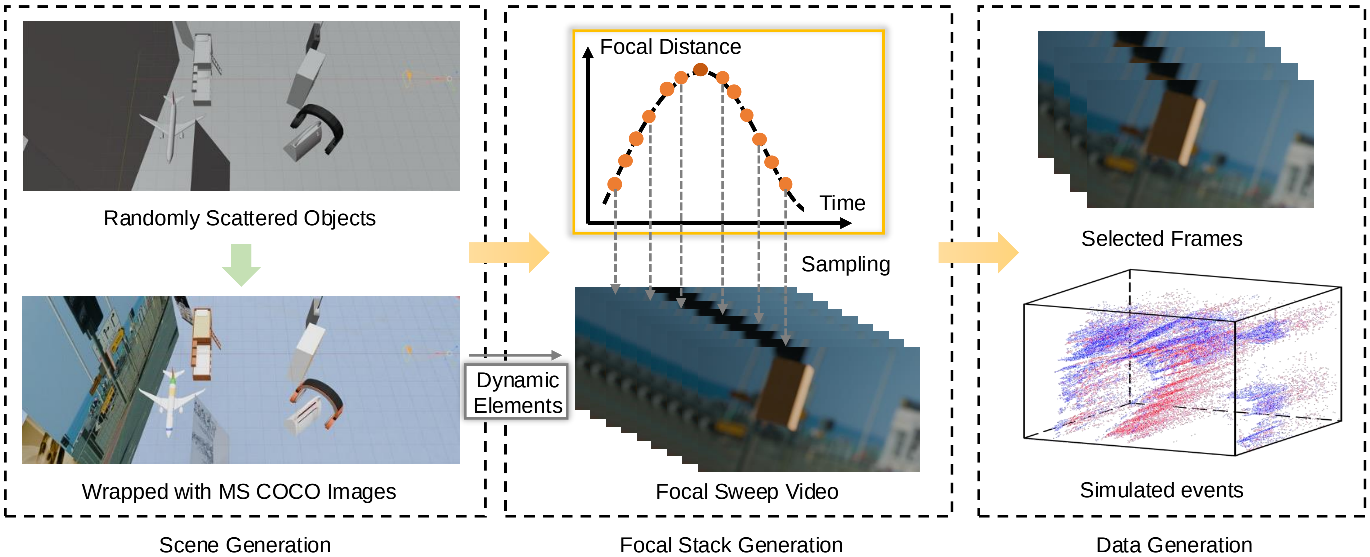


Fig. 1. An illustration of our synthetic data generation pipeline.

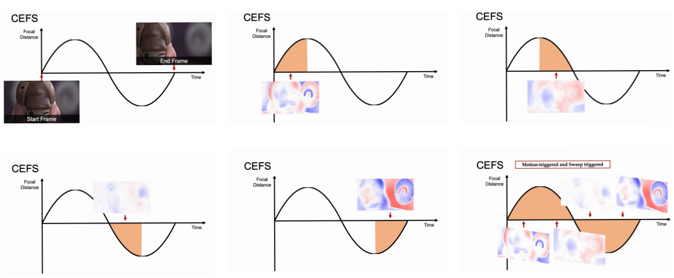


Fig. 2. An illustrative example of the CEFS. The same sequence is provided in the supplementary video.

V. Application

Guided by CEFS, our method supports post-capture focus tracking and all-in-focus (AIF) imaging. Furthermore, by generating intermediate frames, it can also produce AIF videos. Examples of focus tracking, AIF imaging, and AIF video generation are presented in the supplementary video. The results highlight our method's capability to track focus on moving objects across varying focal distances. For AIF imaging, our approach leverages CEFS to guide the warping of texture information to the all-in-focus timestamp, producing sharp and comprehensive AIF images.

VI. More Results on Synthetic Data

In this section, we provide more qualitative comparisons among our method, DRBNet^{+} [5], RefocuGAN [6], and EFS [3] on synthetic data, shown in Figure 3 (`DF-Motion` dataset) and Figure 4 (`DF-Shake` dataset). Compared with image-based methods, our approach restores more high-frequency details encoded within the CEFS. Unlike the EFS method [3], where motion-triggered events introduce artifacts and misalignment in the refocused image, our method avoids these issues by separating the two types of events.

VII. More Results on Real Data

In this section, we provide more qualitative comparisons among our method, DRBNet^{+} [5], RefocuGAN [6], and EFS [3] on real data, shown in Figure 5. Our method effectively recovers refocused images with precise texture in defocused regions, whereas image-based methods fail to restore sharp details. In contrast, EFS [3] suffers from motion-triggered events, leading to additional artifacts.

References

[1] Lin, S., Ma, Y., Guo, Z., Wen, B.: DVS-Voltmeter: Stochastic process-based event simulator for dynamic vision sensors. In: Proc. of European Conference on Computer Vision. pp. 578–593 (2022)
[2] Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Proc. of European Conference on Computer Vision (2014)
[3] Lou, H., Teng, M., Yang, Y., Shi, B.: All-in-focus imaging from event focal stack. In: Proc. of Computer Vision and Pattern Recognition. pp. 17366–17375 (2023)
[4] Peng, J., Cao, Z., Luo, X., Lu, H., Xian, K., Zhang, J.: BokehMe: When neural rendering meets classical rendering. In: Proc. of Computer Vision and Pattern Recognition. pp. 16283–16292 (2022)
[5] Ruan, L., Chen, B., Li, J., Lam, M.: Learning to deblur using light field generated and real defocus images. In: Proc. of Computer Vision and Pattern Recognition. pp. 16304–16313 (2022)
[6] Sakurikar, P., Mehta, I., Balasubramanian, V.N., Narayanan, P.: RefocusGAN: Scene refocusing using a single image. In: Proc. of European Conference on Computer Vision. pp. 497–512 (2018)

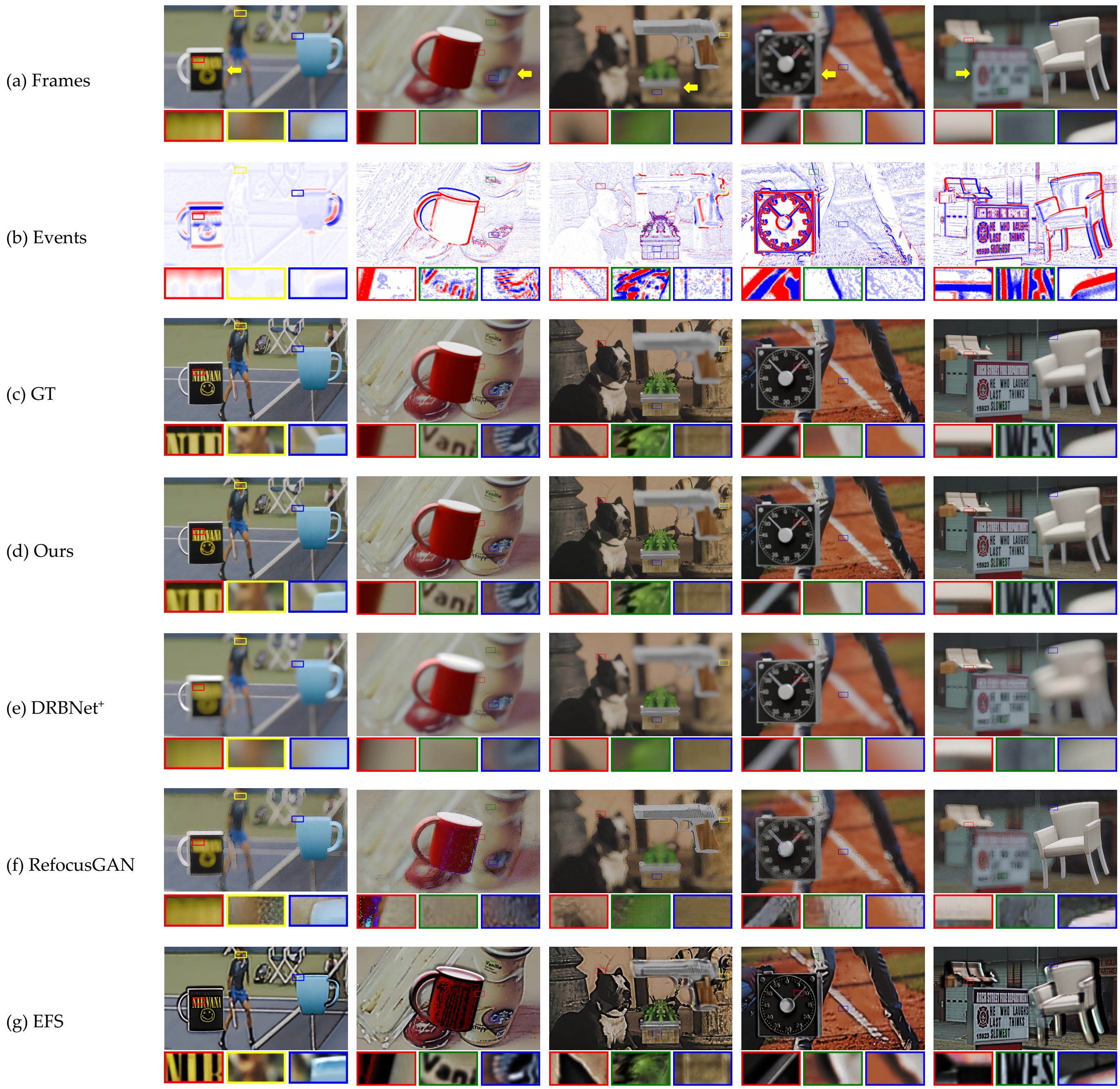


Fig. 3. Visual quality comparison with refocusing methods on the `DF-Motion` dataset. (a) Input frames. (b) Events. (c) Ground truth. (d)∼(g) Refocusing results of ours, DRBNet^{+} [5], RefocusGAN [6], and EFS [3]. Yellow arrows indicate the target focal points. Close-up views are provided below each image.

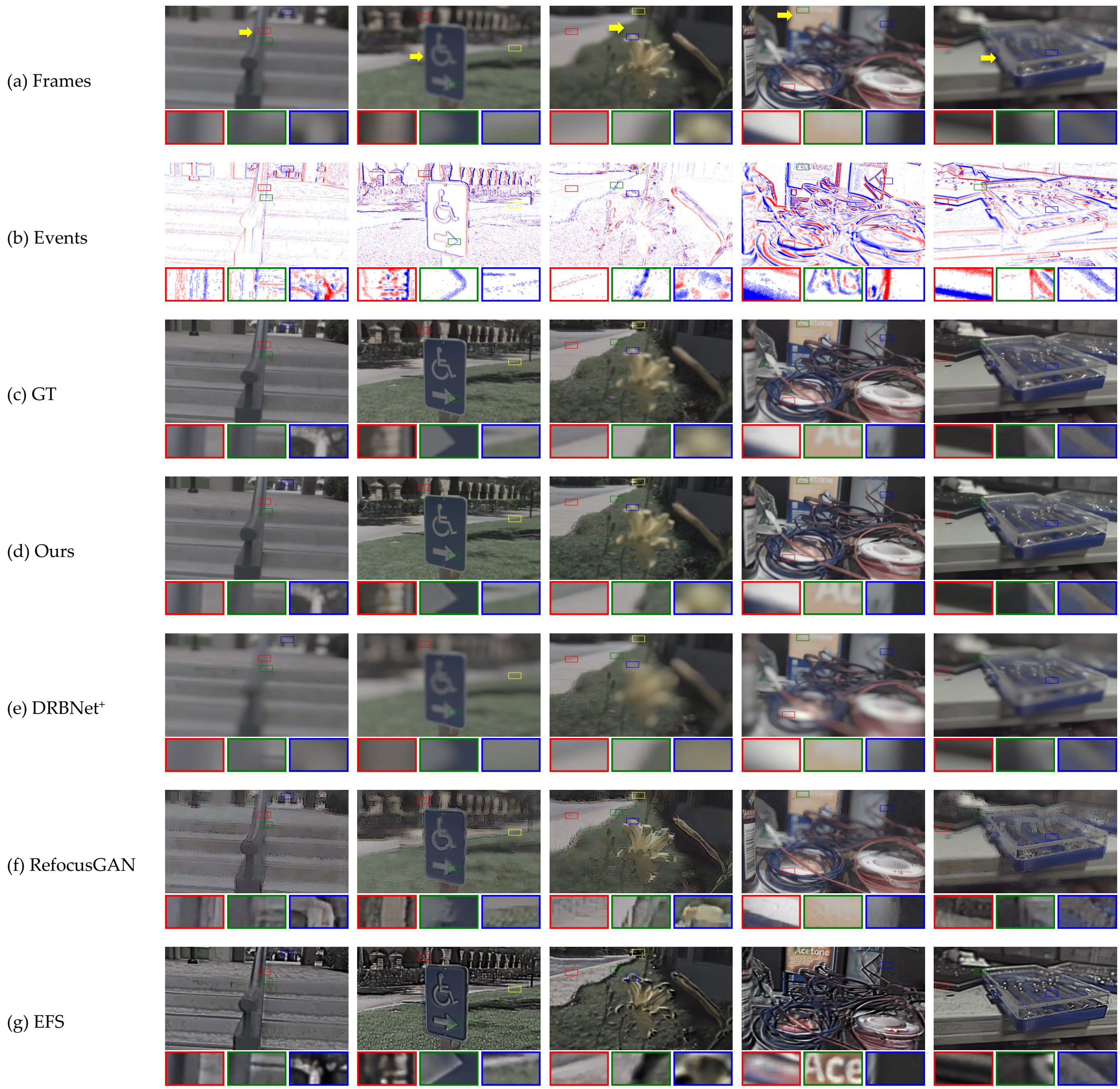


Fig. 4. Visual quality comparison with refocusing methods on the `DF-Shake` dataset. (a) Input frames. (b) Events. (c) Ground truth. (d)∼(g) Refocusing results of ours, DRBNet^{+} [5], RefocusGAN [6], and EFS [3]. Yellow arrows indicate the target focal points. Close-up views are provided below each image.

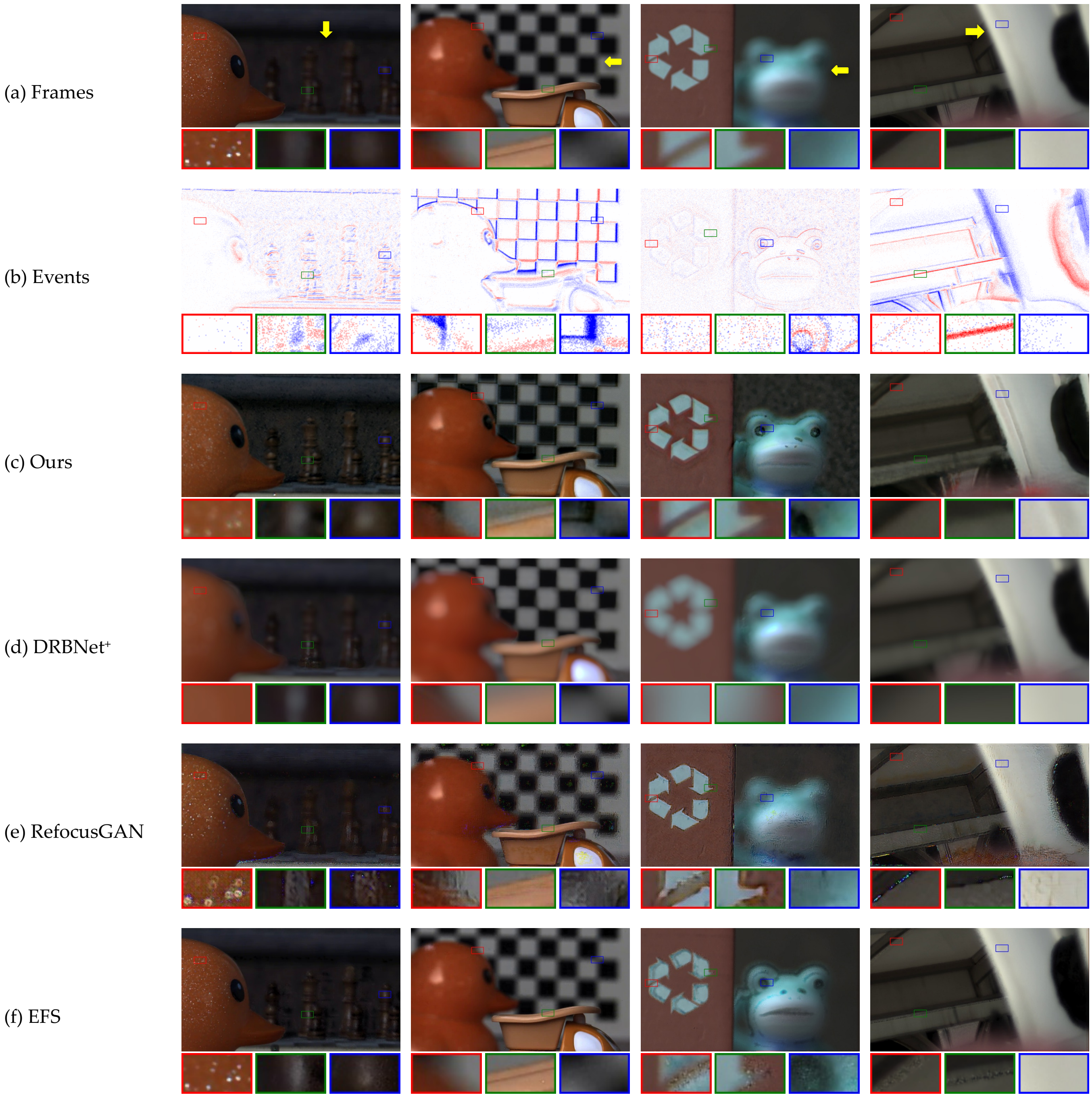


Fig. 5. Visual quality comparison with refocusing methods on the real data. (a) Input frames. (b) Events. (c)∼(f) Refocusing results of ours, DRBNet^{+} [5], RefocusGAN [6], and EFS [3]. Yellow arrows indicate the target focal points. Close-up views are provided below each image.