Autofocus for Event Cameras

Shijie Lin[1], Yinqiang Zhang[1], Lei Yu[2], Bin Zhou[3], Xiaowei Luo[4], Jia Pan[1] *

[1]The University of Hong Kong   [2]Wuhan University   [3]Beihang University   [4]City University of Hong Kong

Project page: https://eleboss.github.io/eaf_webpage/

Abstract

Focus control (FC) is crucial for cameras to capture sharp images in challenging real-world scenarios. The autofocus (AF) facilitates the FC by automatically adjusting the focus settings. However, due to the lack of effective AF methods for the recently introduced event cameras, their FC still relies on naive AF like manual focus adjustments, leading to poor adaptation in challenging real-world conditions. In particular, the inherent differences between event and frame data in terms of sensing modality, noise, temporal resolutions, etc., bring many challenges in designing an effective AF method for event cameras. To address these challenges, we develop a novel event-based autofocus framework consisting of an event-specific focus measure called event rate (ER) and a robust search strategy called event-based golden search (EGS). To verify the performance of our method, we have collected an event-based autofocus dataset (EAD) containing well-synchronized frames, events, and focal positions in a wide variety of challenging scenes with severe lighting and motion conditions. The experiments on this dataset and additional real-world scenarios demonstrated the superiority of our method over state-of-the-art approaches in terms of efficiency and accuracy.

1. Introduction

Recently, a novel neuromorphic vision sensor called event camera [3, 24] has gained growing attention with its significant advantages like high dynamic range (HDR, >130 dB [3]) and low latency (1 us [11]). Thus far, event cameras have been widely adopted in various applications, including robotics and computer vision [9], where the focus control (FC) is essential for reliable perception. Simi-
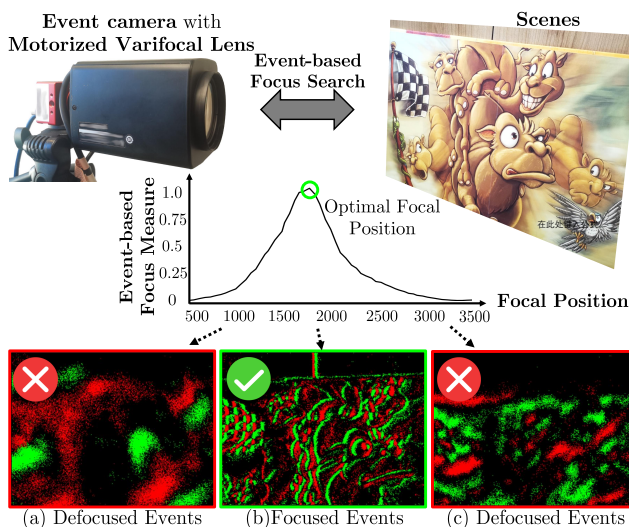
Figure 1. Our event-based autofocus system consists of an event camera and a motorized varifocal lens. It leverages the proposed event-based focus measure and search method to focus the camera to the optimal focal position. When appropriately focused, the event camera's imaging result (b) is sharper and more informative than (a), (c) where it is defocused.

lar to conventional frame-based cameras, the focused events (Fig. 1b) exhibit sharper textures and convey more information than the defocused ones (Fig. 1a and Fig. 1c). Thus, an effective autofocus (AF) method for event cameras is in demand, especially for challenging real-world scenarios.

Conventional AF methods facilitate FC by leveraging the inherent properties of image frames like image gradient [6, 8, 12], image frequency [17, 44], and image statistics [6]. However, events have fundamental differences with frames in terms of modality, noise, temporal resolutions, *etc.*, which make conventional frame-based AF methods not applicable. When developing AF methods specific for event cameras, four particular challenges are:

- **Focus measure function:** Focus measure is a critical function measuring the degree of defocus, but existing frame-based focus measures are designed for 2D image frames but not for the asynchronous event data.

- **Data modality:** Unlike images, events are a stream of four-attribute tuples captured asynchronously with high temporal resolution.

- **Noise:** Events contain noise that are poorly characterised [9] and difficult to be filtered out.

- **Data sizes:** During focusing, event-based AF needs to handle millions of events, making the real-time processing more challenging than conventional AF, which only needs to handle less than one hundred images.

A naive solution to the above challenges is to rely on existing event-based image reconstruction methods [4, 28, 34] to convert events to images and then feed these images to conventional frame-based AF methods. However, the noise of events affects the quality of reconstructed frames and limits the performance of subsequent AF methods. In addition, as a fundamental front-end for numerous event-based applications, AF must be time-efficient, but the learning-based reconstruction is time-consuming. We handle the above challenges by developing an event-based autofocus framework from scratch. First, we developed, according to our knowledge, the first event-based focus measure leveraging the statistics of even rate (ER), which is a simple metric effective for measuring the event data captured at different focal positions. Then, we propose the event-based golden search (EGS) to cooperate with our focus measure to find the optimal focal position. EGS is invariant to the parameter of event accumulation interval and thus can robustly operate in challenging conditions, such as situations with extremely low lighting (<1 Lux) and violent camera shaking. In summary, our contributions are threefold:

- We propose a novel event-based focus measure, *i.e.*, event rate (ER), to measure the focus score of event data at different focal positions. It is efficient, easy to implement, and robust to noise.

- Using our event-based focus measure, we propose a robust and efficient method to focus the event camera by solving a 1-dimensional optimization, which works excellently, especially in complex dynamic and low lighting conditions.

- We collected an event-based autofocus dataset (EAD), containing data under a wide variety of motion and lighting conditions. Extensive evaluations and comparisons have been conducted on EAD and real-world scenarios.

2. Related Work

Frame-based autofocus. Conventional frame-based autofocus methods usually consist of two main components, *i.e.*, the focus measure function and the search method. The focus measure is a unimodal function that scores the imaging data captured at different focal positions. Previous works on focus measure could be categorized into contrast-based, transform-based, statistics-based, and miscellaneous methods. Contrast-based methods [6, 8, 12, 30, 37] leverage the property that focused images are sharper than defocused ones and use the gradient to evaluate the image sharpness. Transform-based methods [7, 17, 19, 22, 23, 38, 42, 44] work in the frequency domain and leverage the energy or frequency ratio to assess the image sharpness. Statistics-based methods [6, 8, 25, 41, 45] attempt to reflect the degree of defocus using statistical characteristics of the image data. Other types of methods include leveraging image curvature [15] and recently introduced learning-based methods [16]. However, these previous works all operate on image inputs in terms of a frame-like tensor. In contrast, the event data is a set of four-attribute spatial-temporal tuples with modality different from that of images. Thus, previous frame-based focus measures cannot be directly exploited to evaluate the focus score for the event camera.

The search methods [14, 18, 21, 40, 43] rely on the focus score computed from images captured at different times to compute the optimal focal position. The global search obtains the optimal result by traversing all possible positions. Later efforts boost the search speed using better sequencing methods, including binary search [18], hill-climbing search [14], Fibonacci search [21, 43], *etc*. Recent work using deep-learning can directly predict the optimal focal position with a few defocused images [40]. However, these approaches were specifically designed for image inputs where each individual image provides rich information in the spatial domain with its large number of pixels though the solution in the temporal domain is low. In contrast, each individual event contains little information, but the set of all asynchronously collected events provides high resolution in the temporal domain though with relatively low resolution in the spatial domain. Thus, novel search methods specific to the event data are necessary for tackling the challenging spatial-temporal imbalance in event data.

Event-based image reconstruction. Frame-based AF algorithms can work on images reconstructed from events. Several recent works [4, 28, 34] demonstrate the possibility of reconstructing intensity images from events. Brandli *et al.* [4] use direct integration to reconstruct frames from events. Munda *et al.* [28] propose a variational model for the event camera and generate events accordingly. But those methods heavily suffer from the noise. The convolutional neural network (CNN) shows outstanding performance [34] in alleviating the noise, but the network's training is difficult to fit the short event accumulation intervals and its execution can be slow on low-cost computational devices, making it not applicable for applications like mobile robots. The contrast maximization framework [10] adopts conventional focus measures as loss functions for maximiz-

ing the projected events, but it cannot solve the autofocus problem. Thus, efficient autofocus methods directly manipulating the event data are still in demand.

3. Problem Formulation

The lens system can be modeled by the *thick lens* model,

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}, \tag{1}$$

where d_o is the object distance, d_i is the image distance, and f is the focal length. In principle, the focused depth is uniquely determined by image-to-object distance $d_i + d_o$ and focal length f, and we sweep the lens to adjust the focal position $p(t)$ toward the optimum $p(t^*)$. As the lens sweep linearly, the image distance and the object distance are varying (with constant velocity) accordingly until Eq. (1) is satisfied, *i.e.*, focused. Let $\epsilon = p(t) - d_i$ be the distance error, defocus happens when the difference between d_i and $p(t)$ is non-zero, i.e. $\epsilon \neq 0$.

Generally, most AF systems follow three main assumptions. **(i) Scenes.** It should have enough textures with relatively high contrast and well-conditioned light environment without high-level flicking. **(ii) Motion.** The field of view should be constrained to the consistent targets during focusing, thus the camera motion cannot be very large. **(iii) Lens.** The lens obeys the *thin lens* model and the motion of the lens motor is linear. During focusing procedure, the AF system continuously moves the lens following a specific searching path and uses time-synchronized data collected along this path as feedback to minimize ϵ. Based on the data type, we can define two AF problems, *i.e.*, the frame-based autofocus (FAF) and event-based autofocus (EAF).

Frame-based autofocus (FAF). FAF uses images taken by frame-based cameras to estimate the optimal focal position:

$$p(t^*) = \underset{t}{\arg\max}\, F_{\mathrm{im}}(S_{\mathrm{im}}(t), t), \tag{2}$$

where F_{im} is the frame-based focus measure function, S_{im} is a function executing the frame-based intensity sampling from a space-time volume of perceived intensity field ($\Omega \times T \in \mathbb{R}^2 \times \mathbb{R}$). The sampling results of S_{im} during focusing is a set of images $\mathcal{I} = \{I_j\}_{j=1}^{N_f}$, where N_f is the number of total images. As most frame-based cameras are of low FPS ($< 30\,\mathrm{FPS}$), the temporal resolution of the image set is relatively low. Thus, FAF only uses a few temporal samples, *i.e.* images, to find the optimal focal position, making FAF heavily rely on each image's rich spatial content. However, as the spatial signals are easily degraded due to motion blur and low lighting, FAF usually fails to work well in challenging motion and lighting conditions.

Event-based autofocus (EAF). Unlike FAF, EAF uses high temporal resolution events generated by event cameras to estimate the optimal focal position:

$$p(t^*) = \underset{t}{\arg\max}\, F_{\mathrm{ev}}(S_{\mathrm{ev}}(t, \Delta t), t), \tag{3}$$

where F_{ev} is the event-based focus measure function, S_{ev} is the event-based intensity sampling function from the perceived intensity field. Since the event cameras only respond to the relative change of intensity, the sampling results of S_{ev} is a set of events $\mathcal{E} = \{e_k : t_k \in [t - \frac{\Delta t}{2}, t + \frac{\Delta t}{2}]\}_{k=1}^{N_e}$, controlled by the sampling time t and sampling interval Δt, and N_e is the number of total events. Generally, N_e is several orders of magnitude larger than N_f because the event camera uses a pixel-wise asynchronous response to the intensity change in a very high temporal resolution. In particular, each event only contains little spatial information, *i.e.*, a pixel's triggering time t_k and polarity p_k. Thus, EAF needs to solve the AF problem by relying on abundant temporal samples while each sample only contains limited spatial information. Since the data modality of events is fundamentally different from that of images, challenges exist when developing efficient EAF algorithms to fully exploit the advantages of event cameras like HDR, low motion blur, and high-temporal resolutions in the EAF task.

4. Methods

In this section, we develop a complete EAF framework from scratch to fully exploit the event data for focusing. The system first traverses all focal positions to collect events (Fig. 2a). Then the proposed event-based golden search (Sec. 4.3) will use our event-based focus measure (Sec. 4.2) to evaluate the collected events and compute the optimal focal position (Fig. 2b and Fig. 2c).

4.1. Preliminaries

From literature [20, 29], we know that defocus is a low-pass filter that attenuates high-frequency components in images. The intensity gradient represents the high-frequency component and can effectively reflect the degree of defocus [13]. Thus, conventional frame-based focus measure function in Eq. (2) can be defined as:

$$F_{\mathrm{im}}(S_{\mathrm{im}}(t), t) \stackrel{\text{def}}{=} \sum_{\boldsymbol{x} \in \Omega} |\nabla I(\boldsymbol{x}, t)|^2, \tag{4}$$

where $\nabla I(\boldsymbol{x}, t) = \left(\frac{\partial I(t)}{\partial x}, \frac{\partial I(t)}{\partial y} \right)^{\top}$ is the gradient at pixel $\boldsymbol{x}$ of an intensity image sampled at time t. Then frame-based AF methods can solve the optimization problem in Eq. (2) to find the focal position. However, for the event camera, the problem of finding an effective event-based focus measure to measure the event data is still open.

(a) Traverse positions from min to max   (b) Searching for the optimal focal position with EGS and ER   (c) Set the lens to optimum
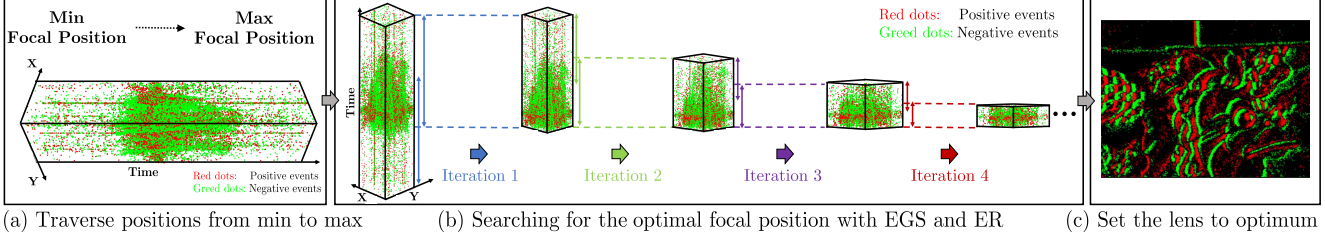
Figure 2. Our event-based autofocus system first traverses all possible focal positions from minimum to maximum to collect events data. (b) Then the system will use the event-based golden search (EGS) in cooperation with the event-based focus measure, *i.e.*, ER, to find the optimal focal position and (c) adjust the lens accordingly.

4.2. Focus Measures for Event cameras

We design an effective event-based focus measure by starting from the basic formation model of event cameras.

4.2.1 Event Formation Model

Event cameras [3,24] asynchronously respond to the brightness change. A pixel in an event camera triggers an event e_k once the change of the log intensity relative to last time step exceeds a predefined contrast threshold C:

$$\Delta L\left(\boldsymbol{x}_k, t_k\right) \stackrel{\text{def}}{=} L\left(\boldsymbol{x}_k, t_k\right) - L\left(\boldsymbol{x}_k, t_k - \Delta t_k\right) = p_k \cdot C, \tag{5}$$

where $\boldsymbol{x}_k$ is the pixel position of the k-th event, t_k is the triggering time, $p_k \in \{-1, +1\}$ is the polarity, Δt_k is the duration since the last triggered event at pixel $\boldsymbol{x}_k$, and $L(\boldsymbol{x}, t) \stackrel{\text{def}}{=} \log I(\boldsymbol{x}, t)$ denotes the log intensity.

The accumulation of events in a duration of t can be computed by the addition of the noise and the difference between log intensities:

$$\int_0^t \sum_k e_k(\boldsymbol{x}, \tau) d\tau = L(\boldsymbol{x}, t) - L(\boldsymbol{x}, 0) + \int_0^t \eta(\boldsymbol{x}, \tau) d\tau, \tag{6}$$

where $\eta(\boldsymbol{x}, t)$ is the sensor noise, which is generally unknown [9] and poorly characterised, and $e_k(\boldsymbol{x}, t)$ is the model of each event using Dirac-delta functions δ [27]:

$$e_k(\boldsymbol{x}, t) \stackrel{\text{def}}{=} p_k \cdot C \cdot \delta\left(t - t_k\right) \delta\left(\boldsymbol{x} - \boldsymbol{x}_k\right). \tag{7}$$

4.2.2 Event-based Focus Measure: Event Rates (ER)

Event-based focus measures aim to efficiently provide a focus score for the events captured at different focal positions. However, classical methods for computing the gradient image [13] need spatial convolution and cannot be directly applied for events. Although we can reconstruct images from events, conducting spatial convolutions on so many reconstructed images are time-consuming. As a solution to this difficulty, [35] renders gradient images by applying a linear
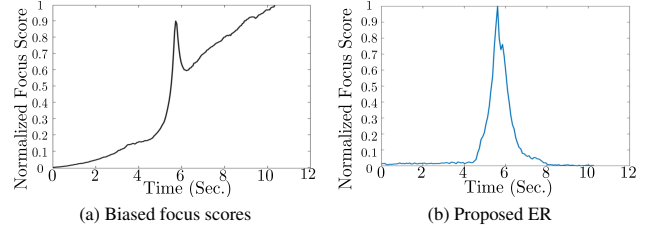

(a) Biased focus scores   (b) Proposed ER

Figure 3. Comparison of focus scores on sequence *focus_board_light_static*. (a) The focus scores using gradient computed by direct integral are affected by the noise accumulated over time, while (b) the focus scores using the event rate is immune from the accumulated noise.

event convolution on both sides of Eq. (6) to skip the step of image reconstruction and directly render gradient images. However, as discussed in [35], computing the direct integral using scheme like [28] will result in drift and biased estimations, as shown in Fig. 3a. Although the drifting can be improved by the linear filter [35], frame-wise summations are still needed to convert gradient images into focus scores and this procedure is time-consuming because the number of events can easily reach millions during focusing.

To address the above issues, we proposed to use event rate (ER) as an event-based focus measure. First, since ER measures the rate of change, the noise will only accumulate in a short interval, largely alleviating the problem of accumulated noise, as shown in Fig. 3b. Second, since the event data are transmitted in an addressed event list from the event camera to computational devices, computing ER can be simplified to count the number of events in the list without generating any frame-like array. In this way, no frame-like rendering or event-wise convolution is needed, leading to a highly efficient ER computation.

We now explain why ER could be an effective event-based focus measure. ER is computed as the average number of events within the event accumulation interval with the length Δt:

$$R_e(\boldsymbol{x}, t, \Delta t) = \frac{\int_{t-\Delta t/2}^{t+\Delta t/2} \sum_k \delta(t - t_k) \delta(\boldsymbol{x} - \boldsymbol{x}_k) d\tau}{\Delta t}. \tag{8}$$

As the intensity increments (Eq. (5)) are caused by the

moving edge [11], given a constant velocity during the interval, the intensity changing rate can be approximated as:

$$\frac{\Delta L(\boldsymbol{x},t)}{\Delta t} \approx -\nabla L(\boldsymbol{x},t)^T \cdot \boldsymbol{v}(\boldsymbol{x}), \qquad (9)$$

where $\frac{\Delta L(\boldsymbol{x},t)}{\Delta t}$ denotes the intensity changing rates, $\nabla L(\boldsymbol{x},t)$ is gradient of log intensity [11], and $\boldsymbol{v}(\boldsymbol{x})$ represents the constant velocity of pixel $\boldsymbol{x}$. By changing the interval of integration from $[0,t]$ to $[t-\Delta t/2, t+\Delta t/2]$ in Eq. (6), we further have:

$$\begin{aligned}
\frac{\Delta L(\boldsymbol{x},t)}{\Delta t} &= \frac{L(\boldsymbol{x},t+\Delta t/2) - L(\boldsymbol{x},t-\Delta t/2)}{\Delta t} \\
&= \frac{\int_{t-\Delta t/2}^{t+\Delta t/2} \sum_k e_k(\boldsymbol{x},\tau)d\tau - \int_{t-\Delta t/2}^{t+\Delta t/2} \eta(\boldsymbol{x},\tau)d\tau}{\Delta t}.
\end{aligned} \quad (10)$$

Omitting the noise term in Eq. (10), and replacing $e_k(\boldsymbol{x},\tau)$ by the event model in Eq. (7), we can see that the absolute value of intensity changing rate can be approximated by the proposed ER times the contrast value:

$$\begin{aligned}
\left| \frac{\Delta L(\boldsymbol{x},t)}{\Delta t} \right| &\approx \frac{C \int_{t-\Delta t/2}^{t+\Delta t/2} \sum_k \delta(t-t_k)\delta(\boldsymbol{x}-\boldsymbol{x}_k)d\tau}{\Delta t} \\
&= C \cdot R_e(\boldsymbol{x},t,\Delta t).
\end{aligned} \quad (11)$$

Combining Eq. (11) and Eq. (9), ER and gradient value are related as follows:

$$R_e(\boldsymbol{x},t,\Delta t) \approx \frac{|\nabla L(\boldsymbol{x},t)| \cdot \boldsymbol{v}(\boldsymbol{x})}{C}. \qquad (12)$$

Assuming a constant velocity $\boldsymbol{v}(\boldsymbol{x})$ in Eq. (12), $R_e(\boldsymbol{x},t)$ is proportional to the norm of intensity gradients $\nabla L(\boldsymbol{x},t)$, which implies that ER can be an effective indicator to the norm of intensity gradient. As mentioned in Sec. 4.1, the focal position is identified by the largest norm of intensity gradients. Then due to the proportional relationship between ER and gradient norm, the focusing location can also be found by the position with the largest ER. Therefore, ER can be an effective event-based focus measure and the focus scores is computed as:

$$F_{\text{ev}}(S_{\text{ev}}(t,\Delta t),t) \overset{\text{def}}{=} \sum_{\boldsymbol{x}\in\Omega} R_e^2(\boldsymbol{x},t,\Delta t). \qquad (13)$$

4.3. Optimization

Given the event-based focus measure in Eq. (13), we can solve the optimization problem in Eq. (3) to find the optimal focal position. During optimization, we need to tune the event accumulation interval Δt when computing ER. We first describe a naive optimization with a manually chosen Δt and then introduce our EAF method that can automatically choose the best Δt and eventually the optimal t^*.

Algorithm 1: Naive EAF with a fixed Δt

Data: accumulation interval time Δt, duration T
Result: Optimal focal position $p(t^*)$
$f_{max} = 0$;
for $t \leftarrow 0$ **to** T **do**
  if $F_{ev}(S_{ev}(t,\Delta t),t) > f_{max}$ **then**
    $f_{max} = F_{ev}(S_{ev}(t,\Delta t),t)$;
    $t^* = t$;

Algorithm 2: Event-based Golden Search (EGS)

Data: threshold μ, golden ratio φ
Result: Optimal focal position $p(t^*)$
$[T_1, T_2] \leftarrow$ time range of total events; $T = T_2 - T_1$;
while $T > \mu$ **do**
  $t_1 = T_1 + \frac{\varphi T}{2}$; $t_2 = T_2 - \frac{\varphi T}{2}$; $\Delta t = \varphi \cdot T$;
  if $F_{ev}(S_{ev}(t_1,\Delta t),t_1) \geq F_{ev}(S_{ev}(t_2,\Delta t),t_2)$
  then
    $T_2 = \varphi \cdot T$;
  else
    $T_1 = (1-\varphi) \cdot T$;
  $T \leftarrow T_2 - T_1$;
  $t^* = (T_2 + T_1)/2$;

4.3.1 Naive EAF with Manually Chosen Δt

According to our ER formulation in Eq. (8), given a specific Δt, we can use Eq. (13) to compute focus scores of different t. Among all results, we treat the t that gives the maximum focus score as the optimal t^*. The overall algorithm is summarized in Algorithm 1. An appropriate Δt is important for this naive solution: a too large Δt will decrease the accuracy of the finally estimated t^* while a too small Δt will generate large spiking noise in ER.

4.3.2 EAF with Automated Δt Search

To make EAF immune from the artifacts of a fixed Δt, we develop the event-based golden search (EGS). EGS can automatically adjust the interval Δt while finding t^*. The overall EGS algorithm is shown in Algorithm 2. Specifically, EGS leverages the golden ratio φ to divide the collected event set into two overlapping intervals and computes the focus score of each interval using Eq. (13). Then the entire procedure is repeated for the active interval with a higher ER. Recursively, the active interval will shrink by the same constant proportion in each step, ultimately leading to an efficient way to progressively reduce the interval locating the optimal t^* with the accuracy bound μ, as demonstrated in Fig. 2b.

Note that it is not necessary to perform event counting

(a) *bottle_drone_light_static* (b) *bottle_drone_light_dynamic*

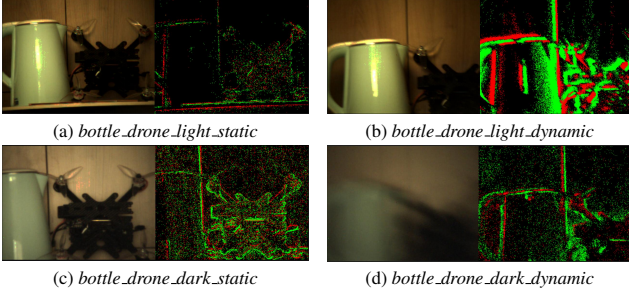(c) *bottle_drone_dark_static* (d) *bottle_drone_dark_dynamic*

Figure 4. Examples of the event-based autofocus dataset. Our dataset contains well-synchronized frames and events in various scenes and conditions. (a), (b), (c), and (d) are captured in *bottle_drone* scene. Please see Appendix for all images in the dataset.

from scratch according to Eq. (8) whenever evaluating event rates, which is computationally redundant and has $\mathcal{O}(N_e^2)$ complexity. Instead, we conduct only once the event counting over the whole event set and record the partial sum results indexed by the event timestamps. In this way, the event rate for any interval can be efficiently computed as the difference of two partial sums indexed by the interval endpoints, resulting in $\mathcal{O}(N_e)$ complexity for EGS.

Even under perfectly static scenes, the lens motion can generate breathing effects, *e.g.*, alternatives between blurriness and sharpness, resulting in brightness change and thus generating events. In this case, the light can spread out over multiple pixels but with attenuation due to the point spread function. Thus, for given light intensity, such attenuation reduces brightness contrast for active pixels, leading to a lower event rate, allowing our method to identify the optimal focal position. Therefore, our method can properly work in both static and dynamic scenes.

5. Experiments

We collect an event-based autofocus dataset (EAD) and compare our methods with several state-of-the-art baselines via a set of qualitative and quantitative experiments. We also conduct multiple tests in various real-world scenes and demonstrate our method's robustness and accuracy.

5.1. The Event-based Autofocus Dataset (EAD)

System setup. As briefly shown in Fig. 1, we equip an event camera (DAVIS346 Color, resolution of 346×260) with the motorized varifocal lens (Fujifilm D17x7.5B-YN1). We use a data acquisition card to capture the position signal of the lens and control the lens from the laptop (Intel i7-1065G7@1.3GHz). The valid focal position ranges from 220 to 3750, which linearly corresponds to the actual focal length from 7.5 mm to 140 mm.

Lighting. We investigate two types of lighting conditions: low lighting (refer to *dark* in EAD) and normal lighting (refer to *light* in EAD). The lowest lighting level can be as low

Table 1. Quantitative comparisons in events. (Better: ↓)

Method	Metric	Scene Type				
		Light		Dark		Total
		Static	Dynamic	Static	Dynamic	
CHEB [45]	MAE	1842.9	1676.7	1806.3	1900.1	1806.5
(recon. image)	RMSE	1931.8	1758.3	1850.0	1947.9	1873.5
HELM [15]	MAE	1387.4	2001.7	1913.9	1900.3	1800.8
(recon. image)	RMSE	1508.8	2067.5	1962.0	1948.1	1883.97
EIGV [41]	MAE	1948.7	1143.4	1906.3	2126.0	1781.1
(recon. image)	RMSE	2040.6	1202.6	1938.4	2146.6	1869.2
GLLV [32]	MAE	1697.1	1634.0	1625.0	1720.7	1669.2
(recon. image)	RMSE	1784.0	1714.4	1726.2	1774.3	1750.0
SML [30]	MAE	1697.1	1634.3	1725.3	1722.7	1694.9
(recon. image)	RMSE	1783.7	1714.6	1779.0	1775.7	1763.5
LAP3 [1]	MAE	1697.0	1634.3	1725.3	1722.7	1694.8
(recon. image)	RMSE	1783.6	1714.6	1779.0	1775.7	1763.5
GRAD [39]	MAE	1697.0	1634.3	1725.3	1722.7	1694.8
(recon. image)	RMSE	1783.6	1714.6	1779.0	1775.7	1763.5
WAVS [44]	MAE	1697.0	1634.1	1725.3	1722.7	1694.8
(recon. image)	RMSE	1783.6	1714.4	1779.0	1775.7	1763.4
ACMO [36]	MAE	1501.9	931.3	1726.0	1828.7	1497.0
(recon. image)	RMSE	1681.0	1105.2	1779.5	1860.4	1633.6
SFIL [26]	MAE	412.9	1633.9	1725.3	1721.0	1373.3
(recon. image)	RMSE	669.7	1714.1	1779.0	1774.5	1557.3
DCTM [22]	MAE	888.3	1078.3	1632.4	1381.6	1245.1
(recon. image)	RMSE	1391.0	1313.2	1758.0	1529.2	1507.3
ER	MAE	98.6	105.3	59.3	87.4	87.6
($\Delta t = 0.055$)	RMSE	112.8	119.2	73.2	113.9	106.4
ER	MAE	91.4	74.3	58.0	67.1	72.7
($\Delta t = 0.065$)	RMSE	107.3	87.6	71.9	103.7	93.7
ER + EGS	MAE	**77.3**	**29.0**	**54.1**	**64.9**	**56.3**
	RMSE	**98.9**	**33.4**	**71.7**	**96.7**	**79.7**

as 0.7 Lux, which is challenging for conventional frame-based cameras and barely visible for human vision.

Motions. We consider two classes of motions that are common in autofocus: shaking motions (introduced by manually shaking the camera during focusing; refer to *dynamic* in EAD) and without shaking motion (refer to *static* in EAD).

Ground-truth. For the ground-truth focal positions, we carefully adjust our system to capture the sharpest image of each scene and record the corresponding focal position. We then verify that this position is also the optimal one for events using high-frequency image reconstruction [31].

We use the setup mentioned above to collect a dataset of frame-event sequences, with sampled snapshots shown in Fig. 4. The dataset includes 28 sequences in four types of combination of lighting and motion conditions, i.e. *static-light*, *static-dark*, *dynamic-light*, and *dynamic-dark*, with 7 sequences for each type. These sequences are representative for a wide variety of real-world scenes, ranging from simple indoor objects to complex outdoor construction sites. Please find the complete statistics of EAD in Appendix.

5.2. Comparisons on Event Data Alone.

Since most event cameras [5, 24, 33] only output events without image frames, it is necessary to complete the EAF task using events alone. In this experiment, we compare our event-based autofocus methods directly working on event data with a set of frame-based autofocus methods [1, 22, 26, 30, 36, 39, 44, 45] operating on 100 FPS images reconstructed from events using [4]. For event-

(a) Focus cruve of sequence: *bottle_drone_light_static*

(b) Focus cruve of sequence: *bottle_drone_light_dynamic*

(c) Focus cruve of sequence: *bottle_drone_dark_static*

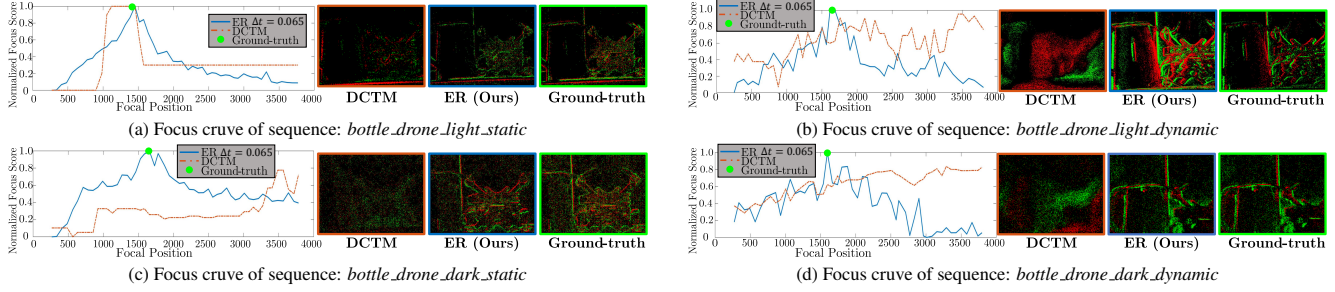(d) Focus cruve of sequence: *bottle_drone_dark_dynamic*

Figure 5. Focus scores in four sequences (a-d) of the proposed event-based focus measure, *i.e.*, the event rate (ER), and the best performed frame-based focus measure in Tab. 1, *i.e.*, DCTM [22]. In each sequence, the sharpest event frame is formed at the optimal focal position (the right image, green edge), in which our ER (the middle image, blue edge) gives the highest response while DCTM (the left image, orange edge) fails to do that.

based autofocus methods, we investigate both the naive solution with different choices of the event accumulation interval (ER with different Δt) and EAF with automated Δt search (ER+EGS). The comparison results are summarized in Tab. 1, where the focusing error is measured by the mean absolute error (MAE) and the root mean square error (RMSE) between the estimated focal position and the ground-truth, averaged over 7 sequences in each combination of lighting and motion conditions. As we can observe, existing frame-based methods cannot handle the noise in events and all of them have much larger MAE and RMSE than event-based methods, indicating their inability to estimate the optimal focal position. In contrast, EAF methods have considerably lower error than frame-based methods, demonstrating their capability to accomplish EAF task in challenging conditions, including large scene dynamics (the *dynamic* class) and low lighting conditions (the *dark* class). Among them, ER+EGS gives the best result in all cases. We also provide the relative error evaluated over each sequence in the Appendix, showing similar results.

ER+EGS's outstanding performance is due to two reasons. First, our proposed ER focus measure can always give highest score around the ground-truth focal positions, as shown in Fig. 5. Such capability to distinguish optimal focal positions is robust, working in both *dynamic* sequences with violent shaking (Fig. 5b and Fig. 5d) and dark indoor scenes with all lights turned off (Fig. 5c and Fig. 5d).

The second reason is EGS algorithm can quickly reduce the focal location error by adaptively tuning the event accumulation interval Δt along with the search procedure. As shown in Fig. 6, we find a fixed stopping threshold $\mu = 0.001$ works well in all challenging conditions with low lighting and violent camera shaking and the resulting ER+EGS is consistently better than the naive search with different values of fixed Δt.

5.3. Comparisons Across Frames and Events

Because most well-established autofocus systems are designed for frame-based cameras, we also conduct cross-

(a) *bottle_drone_light_static*

(b) *bottle_drone_light_dynamic*

(c) *bottle_drone_dark_static*

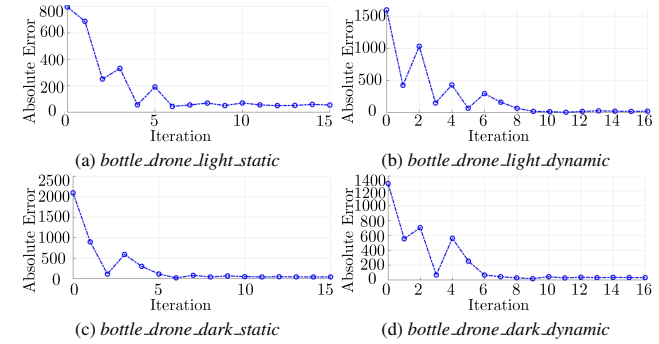(d) *bottle_drone_dark_dynamic*

Figure 6. The curves of absolute error (AE) of focal location for the proposed event-based golden search (EGS). EGS can quickly reduce focal location error on all the four sequences in (a-d) with different lighting and motion conditions.

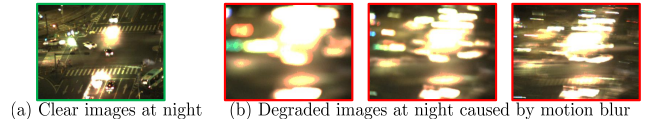(a) Clear images at night   (b) Degraded images at night caused by motion blur

Figure 7. (a) Clear images are easily (b) blurred by shaking.

comparisons on both frames and events. We let conventional frame-based autofocus algorithms run on color images captured by a DAVIS346 camera during the linear focusing. Our EAF methods run on time-synchronized events captured by the same DAVIS346 camera. We report the error between the focal position with the highest focus score and the ground-truth in Tab. 2. As can be observed, conventional frame-based methods work well in static scenes under proper lighting (the *light* + *static* class). However, as dynamics increases, the error of frame-based methods increase, especially in low light dynamic conditions (the *dark* + *dynamic* class) where the motions blur as shown in Fig. 7 results in large errors of conventional methods. In contrast, our EAF method gives accurate estimations in all three challenging conditions, and the proposed ER+EGS achieves the lowest error except in static scenes with normal lighting.

We also notice that some frame-based EAF methods out-

Table 2. Cross-comparisons in events and frames. (Better: ↓)

Method	Metric	Scene Type				
		Light		Dark		Total
		Static	Dynamic	Static	Dynamic	
ACMO [36]	MAE	1431.7	1382.9	1880.9	1726.7	1605.5
(color image)	RMSE	1573.7	1517.5	1931.3	1791.1	1711.5
DCTM [22]	MAE	432.1	594.0	1215.1	1630.1	967.9
(color image)	RMSE	579.5	780.8	1448.7	1680.5	1211.2
EIGV [41]	MAE	690.4	658.1	951.3	1297.6	899.4
(color image)	RMSE	1010.7	936.7	1217.9	1420.4	1161.8
CHEB [45]	MAE	**37.3**	347.6	499.1	1425.6	577.4
(color image)	RMSE	**50.6**	792.7	781.6	1587.8	969.9
SML [30]	MAE	44.0	51.1	483.9	1380.0	489.8
(color image)	RMSE	52.0	71.7	787.2	1628.4	905.4
WAVS [44]	MAE	44.0	56.3	405.1	1060.7	391.5
(color image)	RMSE	52.0	77.2	768.3	1415.8	806.8
LAP3 [1]	MAE	41.6	53.0	124.0	1060.9	319.9
(color image)	RMSE	48.8	75.7	165.6	1416.0	714.3
SFIL [26]	MAE	41.3	104.4	124.0	765.7	258.9
(color image)	RMSE	63.7	147.3	165.6	1070.4	547.5
GLLV [32]	MAE	46.6	38.9	191.7	748.6	256.4
(color image)	RMSE	68.7	52.7	266.3	1085.7	560.6
HELM [15]	MAE	38.4	47.1	124.0	772.1	245.4
(color image)	RMSE	51.2	70.4	165.6	1087.5	551.7
GRAD [39]	MAE	41.1	57.0	124.0	421.4	160.9
(color image)	RMSE	49.3	72.0	165.6	854.4	437.3
ER	MAE	98.6	105.3	59.3	87.4	87.6
($\Delta t = 0.055$)	RMSE	112.8	119.2	73.2	113.9	106.4
ER	MAE	91.4	74.3	58.0	67.1	72.7
($\Delta t = 0.065$)	RMSE	107.3	87.6	71.9	103.7	93.7
ER + EGS	MAE	77.3	**29.0**	**54.1**	**64.9**	**56.3**
	RMSE	98.9	**33.4**	**71.7**	**96.7**	**79.7**

(a) Dynamic scenes: the fan is working, and the turntable is rotating the box.

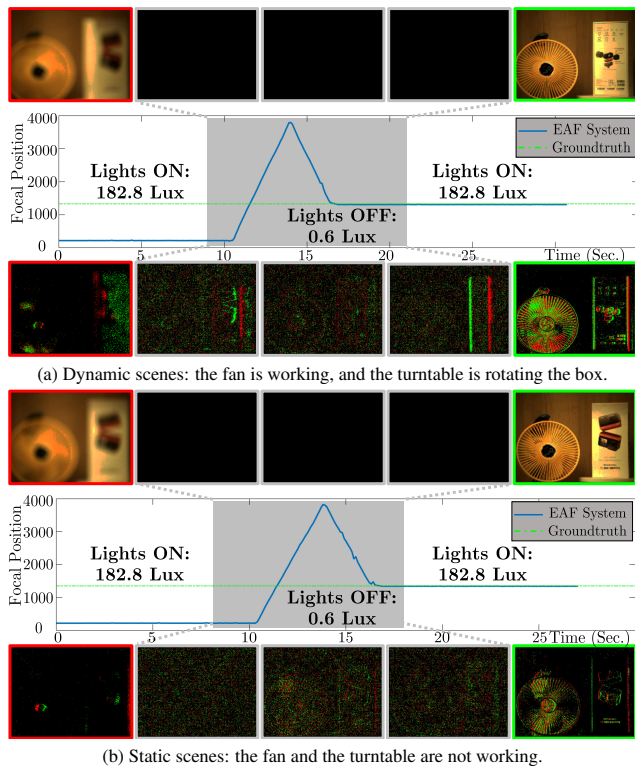(b) Static scenes: the fan and the turntable are not working.

Figure 8. From left to right, our system accurately focus the camera in low light (marked by gray) in two situations with and without scenes motions.

perform our method in relatively simple situations that are static and with good lighting conditions (*static+light* column in Tab. 2). The reason is a combination of noise and the simplicity of the scenes. In static scenes, events are generated by lens motions. Thus, once the scenes contain too many low-contrast textures, the activated events will be limited, which might be even reduced since the point spread function will attenuate the texture contrast. As such, the insufficient number of events will bring more bias to find the optimal focusing point since events are always contaminated by noise. Our EGS divides the total events into two intervals, and each contains an even smaller number of events, thus with lower signal-to-noise ratio (SNR), leading to sub-optimal results, increasing the ultimately averaged errors. Besides, we further provide relative error evaluated over each sequence in the Appendix. Interested readers can refer to it for detailed inspection.

5.4. EAF in extremely low lighting conditions

To further illustrate the robustness of our EAF system, we perform additional AF experiments in extremely low lighting conditions, by turning the light off before the focusing procedure starts (corresponding to the gray area in Fig. 8a and Fig. 8b) in the room, and then turn the light on after the focusing finishes to check the quality of the optimal focus position estimated in dark. Results are shown in Fig. 8, demonstrating that our EAF system can sharply focus images and events in a very dark room ($< 0.7 \, \text{Lux}$),

while due to the low dynamic range of frame-based parts in the DAVIS camera, images captured in the low light are heavily degraded and cannot support the AF tasks. More details are in the video in supplemental materials.

6. Conclusion and Discussion

In this paper, we provide a novel solution to the event-based autofocus task, consisting of a simple and effective focus measure called event rate (ER) and a robust event-based golden search (EGS) to efficiently find the optimal focal position for sharp imaging. We also collected the EAF dataset (EAD) with well-synchronized frames, events, and focal positions in a wide variety of motion and lighting conditions. Extensive experiments have verified the accuracy of our method when focusing an event-based camera in challenging conditions with low lighting, violent camera shaking, and complex background.

Discussion: We found that the events usually suffer from the temporal delay under higher-contrast changes. This delay breaks the synchronization and increases the focusing error. In the future, we will exploit methods like inceptive events [2] to address this sensing limitation.

References

[1] Youngeun An, Gwangwon Kang, Il-Jung Kim, Hyun-Sook Chung, and Jongan Park. Shape from focus through laplacian using 3d window. In *2008 Second International Conference on Future Generation Communication and Networking*, volume 2, pages 46–50, 2008. 6, 8

[2] R Baldwin, Mohammed Almatrafi, Jason R Kaufman, Vijayan Asari, and Keigo Hirakawa. Inceptive event time-surfaces for object classification using neuromorphic cameras. In *International Conference on Image Analysis and Recognition (ICIAR)*, pages 395–403. Springer, 2019. 8

[3] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240×180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 1, 4

[4] Christian Brandli, Lorenz Muller, and Tobi Delbruck. Real-time, high-speed video decompression using a frame-and event-based davis sensor. In *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 686–689, 2014. 2, 6

[5] Shoushun Chen and Menghan Guo. Live demonstration: Celex-v: a 1m pixel multi-mode event-based sensor. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1682–1683, 2019. 6

[6] N Ng Kuang Chern, Poo Aun Neow, and Marcelo H Ang. Practical issues in pixel-based autofocusing for machine vision. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, volume 3, pages 2791–2796, 2001. 1, 2

[7] Kanjar De and V Masilamani. Image sharpness measure for blurred images in frequency domain. *Procedia Engineering*, 64:149–158, 2013. 2

[8] Lawrence Firestone, Kitty Cook, Kevin Culp, Neil Talsania, and Kendall Preston Jr. Comparison of autofocus methods for automated microscopy. *Cytometry: The Journal of the International Society for Analytical Cytology*, 12(3):195–206, 1991. 1, 2

[9] Guillermo Gallego, Tobi Delbruck, Garrick Michael Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 1, 2, 4

[10] Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza. Focus is all you need: Loss functions for event-based vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12280–12289, 2019. 2

[11] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Eklt: Asynchronous photometric feature tracking using events and frames. *International Journal of Computer Vision*, 128(3):601–618, 2020. 1, 5

[12] Jan-Mark Geusebroek, Frans Cornelissen, Arnold WM Smeulders, and Hugo Geerts. Robust autofocusing in microscopy. *Cytometry: The Journal of the International Society for Analytical Cytology*, 39(1):1–9, 2000. 1, 2

[13] Rafael C Gonzalez and Richard E Woods. Digital image processing, hoboken, 2018. 3, 4

[14] Jie He, Rongzhen Zhou, and Zhiliang Hong. Modified fast climbing search auto-focus algorithm with adaptive step size searching technique for digital camera. *IEEE transactions on Consumer Electronics*, 49(2):257–262, 2003. 2

[15] Franz Stephan Helmli and Stefan Scherer. Adaptive shape from focus with an error estimation in light microscopy. In *Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis*, pages 188–193, 2001. 2, 6, 8

[16] Charles Herrmann, Richard Strong Bowen, Neal Wadhwa, Rahul Garg, Qiurui He, Jonathan T Barron, and Ramin Zabih. Learning to autofocus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2230–2239, 2020. 2

[17] Jaroslav Kautsky, Jan Flusser, Barbara Zitova, and Stanislava Šimberová. A new wavelet-based measure of image focus. *Pattern recognition letters*, 23(14):1785–1794, 2002. 1, 2

[18] Nasser Kehtarnavaz and H-J Oh. Development and real-time implementation of a rule-based auto-focus algorithm. *Real-Time Imaging*, 9(3):197–203, 2003. 2

[19] Matej Kristan, Janez Perš, Matej Perše, and Stanislav Kovačič. A bayes-spectral-entropy-based measure of camera focus using a discrete cosine transform. *Pattern Recognition Letters*, 27(13):1431–1439, 2006. 2

[20] Eric Krotkov. Focusing. *International Journal of Computer Vision*, 1:223–237, 1988. 3

[21] Eric P Krotkov. *Active computer vision by cooperative focus and stereo*. Springer Science & Business Media, 2012. 2

[22] Sang-Yong Lee, Yogendera Kumar, Ji-Man Cho, Sang-Won Lee, and Soo-Won Kim. Enhanced autofocus algorithm using robust focus measure and fuzzy reasoning. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(9):1237–1246, 2008. 2, 6, 7, 8

[23] Sang-Yong Lee, Jae-Tack Yoo, Yogendera Kumar, and Soo-Won Kim. Reduced energy-ratio measure for robust auto-focusing in digital camera. *IEEE Signal Processing Letters*, 16(2):133–136, 2009. 2

[24] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15μs latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2):566–576, 2008. 1, 4, 6

[25] Shuxin Liu, Manhua Liu, and Zhongyuan Yang. An image auto-focusing algorithm for industrial image measurement. *EURASIP Journal on Advances in Signal Processing*, 2016(1):1–16, 2016. 2

[26] Rashid Minhas, Abdul A Mohammed, QM Jonathan Wu, and Maher A Sid-Ahmed. 3d shape from focus and depth map computation using steerable filters. In *International conference image analysis and recognition*, pages 573–583. Springer, 2009. 6, 8

[27] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, 2017. 4

[28] 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. 2, 4

[29] Shree K Nayar and Yasuo Nakagawa. Shape from focus: An effective approach for rough surfaces. In *Proceedings., IEEE International Conference on Robotics and Automation (ICRA)*, pages 218–225, 1990. 3

[30] Shree K Nayar and Yasuo Nakagawa. Shape from focus. *IEEE Transactions on Pattern analysis and machine intelligence*, 16(8):824–831, 1994. 2, 6, 8

[31] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6813–6822, June 2019. 6

[32] José Luis Pech-Pacheco, Gabriel Cristóbal, Jesús Chamorro-Martinez, and Joaquín Fernández-Valdivia. Diatom autofocusing in brightfield microscopy: a comparative study. In *Proceedings 15th International Conference on Pattern Recognition (ICPR)*, volume 3, pages 314–317, 2000. 6, 8

[33] Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems*, 33, 2020. 6

[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*, 2019. 2

[35] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Asynchronous spatial image convolutions for event cameras. *IEEE Robotics and Automation Letters*, 4(2):816–822, 2019. 4

[36] Mukul V Shirvaikar. An optimal measure for camera focus and exposure. In *Thirty-Sixth Southeastern Symposium on System Theory, 2004. Proceedings of the*, pages 472–475, 2004. 6, 8

[37] Murali Subbarao, Tae-Sun Choi, and Arman Nikzad. Focusing techniques. *Optical Engineering*, 32(11):2824–2836, 1993. 2

[38] Ran Tao, Wei Zhang, and Yanlei Li. Time–frequency filtering-based autofocus. *Signal processing*, 91(6):1401–1408, 2011. 2

[39] Jay Martin Tenenbaum. *Accommodation in computer vision*. Stanford University, 1971. 6, 8

[40] Chengyu Wang, Qian Huang, Ming Cheng, Zhan Ma, and David J Brady. Intelligent autofocus. *arXiv preprint arXiv:2002.12389*, 2020. 2

[41] Chong-Yaw Wee and Raveendran Paramesran. Measure of image sharpness using eigenvalues. *Information Sciences*, 177(12):2533–2552, 2007. 2, 6, 8

[42] Hui Xie, Weibin Rong, and Lining Sun. Wavelet-based focus measure and 3-d surface reconstruction method for microscopy images. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 229–234, 2006. 2

[43] Yalin Xiong and Steven A Shafer. Depth from focusing and defocusing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (ICCV)*, pages 68–73, 1993. 2

[44] Ge Yang and Bradley J Nelson. Wavelet-based autofocusing and unsupervised segmentation of microscopic images. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 3, pages 2143–2148, 2003. 1, 2, 6, 8

[45] Pew Thian Yap and P Raveendran. Image focus measure based on chebyshev moments. *IEE Proceedings-Vision, Image and Signal Processing*, 151(2):128–136, 2004. 2, 6, 8