One-Step Event-Driven High-Speed Autofocus

Yuhan Bao Shaohua Gao Wenyong Li Kaiwei Wang*
State Key Laboratory of Extreme Photonics and Instrumentation, Zhejiang University, China
{yhbao, gaoshaohua, liwenyong2023, wangkaiwei}@zju.edu.cn

Abstract

High-speed autofocus in extreme scenes remains a significant challenge. Traditional methods rely on repeated sampling around the focus position, resulting in "focus hunting". Event-driven methods have advanced focusing speed and improved performance in low-light conditions; however, current approaches still require at least one lengthy round of "focus hunting", involving the collection of a complete focus stack. We introduce the Event Laplacian Product (ELP) focus detection function, which combines event data with grayscale Laplacian information, redefining focus search as a detection task. This innovation enables the first one-step event-driven autofocus, cutting focusing time by up to two-thirds and reducing focusing error by 24 times on the DAVIS346 dataset and 22 times on the EVK4 dataset. Additionally, we present an autofocus pipeline tailored for event-only cameras, achieving accurate results across a range of challenging motion and lighting conditions. All datasets and code are available in https://github.com/YuHanBaozju/ELP.

1. Introduction

Focus is a prerequisite for most visual tasks. An ideal autofocus (AF) system guides the motor directly toward the correct focus position from any starting point, and stop immediately upon arrival, ensuring precision and efficiency, which can be considered as "one-step AF".

Conventional AF algorithms predict focus by evaluating the image contrast [19] (or sharpness) at different points, which often requires repeated sampling around the focus position, leading to "focus hunting". At the same time, the contrast-based AF methods are susceptible to low frame rates, motion blur. To enhance speed and accuracy, Phase Detection AutoFocus (PDAF) with dual-pixel sensors is commonly employed. These sensors measure phase differences to calculate focus position. However, PDAF is confronted with several challenges: the complexity of pixel design limits the number of dual pixels, reduced light intake impacts performance in low-light conditions, and it struggles in scenarios with significant defocus.

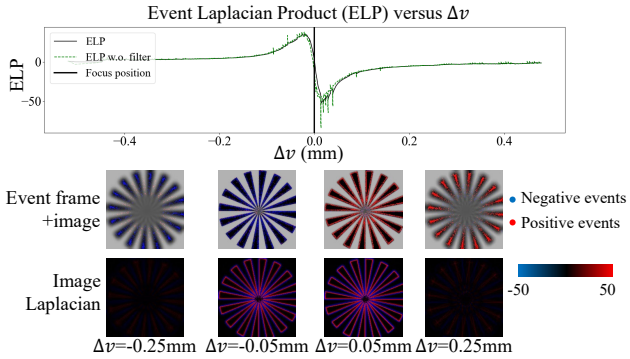


Figure 1. Visualization of ELP, event frames, images and Laplacian at different Δv.

Recent research has shown the potential of event cameras for high-speed AF applications. Event cameras detect brightness changes at individual pixels with extreme temporal resolution, down to 1 μs, making them ideal for dynamic scenarios. However, the unique asynchronous characteristics of event cameras set event-driven AF methods apart from traditional frame-based approaches. The use of event rate (ER) as a focus evaluation function was first proposed in [10], where the focus position is identified by locating the position with the highest ER in the focus stack. Building on this, the Event Golden Search (EGS) algorithm was introduced to improve the speed of focus search. Additionally, another study [1] noted that brightness changes exhibit symmetry around the focus position in the focus stack. Building on this insight, the Polarity Based Autofocus (PBF) algorithm was developed, effectively leveraging this symmetry to achieve rapid and precise focusing across various lighting conditions, including strobe lighting. However, current event-driven AF algorithms require analyzing a complete focus stack, including information both before and after the focus position, before the focus position can be determined. While these methods eliminate repeated "focus hunting", they still require a single round of it, which involves capturing the entire focus stack, searching for the focus position, and moving there.

For the first time, we integrate both events and grayscale information to achieve one-step event-driven AF. We theoretically derive a fundamental connection between the spatial second-order derivative and the temporal first-order derivative of the image during the focusing process. Building on this, we introduce a focus **detection** function called the "Event Laplacian Product" (ELP), which shows a distinct "sign mutation" at the focus position, as shown in Fig. 1. The proposed ELP method determines the focus position in real time by detecting the "sign mutation" of the ELP value from positive to negative, eliminating the need for iterative searches required by traditional methods based on the "peaked" focus evaluation function. Additionally, ELP predicts the direction of focus adjustment based on the sign of its value, effectively serving as a focus **prediction** function. Moreover, by integrating ELP with the latest event-based temporal mapping photography [2], we achieve one-step high-speed AF using event only.

To the best of our knowledge, we are the first to implement event-driven one-step AF, achieving accurate focusing with less than one depth of focus across various brightness conditions and motion states in the synthetic, DAVIS346, and Prophesee EVK4 datasets. Compared to existing state-of-the-art event-driven AF method, ELP reduces focusing error by **24 times** on the DAVIS346 dataset and by **22 times** on the EVK4 dataset. Additionally, the focusing time is reduced by **two-thirds**. Compared to the current one-step AF method PDAF, our approach is not limited by the number of dual-pixels, remains robust in low-light conditions, and performs reliably even under significant defocus.

2. Related Work

2.1. Conventional camera AF

Currently, the widely-used AF methods for conventional cameras are contrast-based AF and PDAF, or a mixture of both. The contrast-based AF algorithm consists of two key components: the focus evaluation function and the search algorithm. The focus evaluation function typically includes: (1) First-order gradient methods, such as the Prewitt operator [11] and Sobel operator [13], (2) Second-order gradient methods, like the Laplacian operator [8], and (3) Histogram-based methods [5], among others. Search algorithms commonly used include hill climbing [6] and Fibonacci search [9]. More recently, deep learning-based AF approaches have shown promising results [14], offering some ability to predict the focus position. Yang [18] introduced Differential Focus Volume into the depth from focus field, where the concept of differentiation also serves as an inspiration for AF tasks. However, a significant issue with contrast-based AF is the ambiguity between pre-focus and post-focus, which has been addressed by the introduction of PDAF [16]. By analyzing phase differences between the two elements of dual-pixel sensors, PDAF can determine both the direction and the distance needed for focusing, enabling one-step AF. Nevertheless, the complexity of dual-pixel design and its impact on image quality limit the number of dual-pixel sensors in digital cameras. In conventional camera AF, the frame rate for acquiring raw image data has long been a limiting factor for AF speed. Furthermore, the problem of "focus hunting" in complex scenes—such as low light, motion, and significant defocus—remains a persistent challenge.

2.2. Event-driven AF

The high temporal resolution, dynamic range, and low data redundancy of event cameras offer new possibilities for AF and related applications [15, 17]. The use of ER as a focus evaluation function for event cameras was first introduced in [10], along with the EGS method for quickly determining the focus position within the entire focus stack. However, its focusing principle relies on the assumption that optical flow exists and remains constant, which does not hold in most focusing scenarios. An in-depth investigation of event-driven AF in [1] revealed that pixel brightness changes during focusing exhibit symmetry around the focus position. This insight led to the development of the PBF algorithm, which identifies the focus position by locating the center of symmetry in the event polarity rate (EPR) across the focus stack. Experiments demonstrated the PBF algorithm's robustness in complex scenarios, including both dynamic and static scenes, as well as challenging conditions such as strobe flashes.

Both EGS and PSF, as well as other event-driven methods such as Ge's [4] and Lou's [12], require capturing a complete focus stack, spanning from defocus to near-focus and back to defocus. They analyze the entire stack to locate the focus position and drive the motor to the target position, a process that is often too slow and results in a poor user experience due to "focus hunting". Additionally, the focus evaluation functions they use are "peaked" functions, making them vulnerable to multiple peaks in the presence of interference, which can lead to suboptimal focusing. The most critical limitation of EGS and PBF is their inability to predict in real time whether the focus motor is moving toward the focus position, which can result in misguided "focus hunting", further increasing the focusing time.

3. Proposed method

Our goal is to develop a one-step event-driven AF method that can predict in real time whether to move toward the focus position and promptly signal the focus motor to stop once the focus position is achieved. Compared to existing event-driven AF methods, this approach reduces focusing time by up to two-thirds, including motor runtime. Furthermore, our one-step AF approach eliminates "focus hunt-

ing", significantly streamlining the focusing process.

3.1. Background and Basics

Event camera. Each pixel on the event cameras can respond to changes in the brightness $L(x, y, t)$. Eq. (1) shows how the event camera works:

$$p(x,y,t) = \begin{cases} +1, & if \quad \Delta L > C, \\ -1, & if \quad \Delta L < -C, \end{cases} \tag{1}$$

where $\Delta L = L(x, y, t_i) - L(x, y, t_{i-1})$ denotes the change in brightness since the last trigger event, C denotes the contrast threshold, and $p(x, y, t)$ denotes the event polarity. When the brightness change exceeds the threshold, event cameras emit events with different polarities depending on the direction of the change. Brightness is a logarithmic mappping of light intensity [3]. While events capture information about brightness changes, they do not provide exact intensity values like grayscale images.

Event-driven Temporal-mapping Photography. While passively generated events do not directly represent grayscale information, prior work [2] has demonstrated that the timestamps of events obtained through active transmittance modulation can be mapped to grayscale values. In this method, a motorized aperture modulates the incoming light, allowing the event camera to capture the Initial Positive Event (IPE) for each pixel as the aperture opens. By mapping IPE timestamps to grayscale values, this method enables accurate high-dynamic-range (HDR) grayscale imaging. The detailed mapping relationship is provided in Eq. (4) of [2]. Event-driven Temporal-Mapping photography (EvTemMap) provides valuable grayscale information for event-only cameras, which can be utilized in event-driven one-step AF.

3.2. Principle

Principle of focusing optics. For a thin lens, the following equation holds when the image is in focus: $\frac{1}{u} + \frac{1}{v} = \frac{1}{f}$,where u represents the object distance, f represents the focal length of the thin lens, and v represents the ideal image distance. Ideally, the point spread function (PSF) of an optical system in defocus can be modeled as a Gaussian function of the amount of defocus Δv [1]: $h(\boldsymbol{x}, \Delta v) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{\boldsymbol{x}^2}{2\sigma^2}), \sigma = k\frac{|\Delta v|}{2v_0}D$, where k represents the inverse of the sensor pixel size, while v_0 and D represent the ideal image distance and the exit pupil diameter, respectively. And σ can be interpreted as the blur kernel size in pixels. In a real optical system, the PSFs are slightly distorted by aberrations, but their Root Mean Square (RMS) radius always increases with Δv. In a focusing task, the focus motor typically adjusts the image distance v to achieve the position that minimizes $|\Delta v|$, *i.e.*, the point where the blur kernel radius is minimized.

Derivation of image differentiation during the focusing process. Let $F(\boldsymbol{x}, t)$ represent a clear dynamic scene and $h(\boldsymbol{x}, t)$ a Gaussian kernel with its variance σ^2 changing over time during focusing. The image on the image plane of an event camera, $G(\boldsymbol{x}, t)$, is expressed as: $G(\boldsymbol{x}, t) = F(\boldsymbol{x}, t) * h(\boldsymbol{x}, t)$, where $*$ represents convolution.

We can compute the first-order derivative of $G(\boldsymbol{x}, t)$ with respect to time as follows:

$$\frac{\partial}{\partial t}G(\boldsymbol{x},t) = \left(\frac{\partial}{\partial t}F(\boldsymbol{x},t)\right) * h(\boldsymbol{x},t) + F(\boldsymbol{x},t) * \left(\frac{\partial}{\partial t}h(\boldsymbol{x},t)\right), \tag{2}$$

where the first term corresponds to the temporal scene variation and the second term is associated with the focusing process. In the focusing task, the change in $F(t)$ is significantly slower than that in $h(t)$. Therefore, in our derivation, we assume $\frac{\partial}{\partial t}F(\boldsymbol{x}, t) = 0$. For simplicity, assume that the Gaussian variance satisfies $\sigma(t)^2 = \sigma_0^2 + \alpha t$. The expression in Eq. (2) can then be articulated as:

$$\frac{\partial h(\boldsymbol{x},t)}{\partial t} = \frac{\alpha}{2}h(\boldsymbol{x},t)\left(\frac{\boldsymbol{x}^2}{\sigma(t)^4} - \frac{1}{\sigma(t)^2}\right). \tag{3}$$

Similarly, we can compute the spatial second-order derivative of $G(\boldsymbol{x}, t)$. Due to the interchangeability of convolution and differentiation operations, $F(\boldsymbol{x}, t) * \frac{\partial^2 h(\boldsymbol{x},t)}{\partial \boldsymbol{x}^2} = \frac{\partial^2 F(\boldsymbol{x},t)}{\partial \boldsymbol{x}^2} * h(\boldsymbol{x}, t)$. Therefore,

$$\frac{\partial^2 G(\boldsymbol{x},t)}{\partial \boldsymbol{x}^2} = 2F(\boldsymbol{x},t) * \left(\frac{\partial^2 h(\boldsymbol{x},t)}{\partial \boldsymbol{x}^2}\right), \tag{4}$$

where

$$\frac{\partial^2 h(\boldsymbol{x},t)}{\partial \boldsymbol{x}^2} = h(\boldsymbol{x},t)\left(\frac{\boldsymbol{x}^2}{\sigma(t)^4} - \frac{1}{\sigma(t)^2}\right). \tag{5}$$

We can derive $S(t)$ from Eq. (2) and Eq. (4):

$$\begin{aligned} S(t) &= -\int \frac{\partial G(\boldsymbol{x},t)}{\partial t} \cdot \frac{\partial^2 G(\boldsymbol{x},t)}{\partial \boldsymbol{x}^2} \, d\boldsymbol{x} \\ &= -\alpha\left[\int \left(F(\boldsymbol{x},t) * \frac{\partial^2 h(\boldsymbol{x},t)}{\partial \boldsymbol{x}^2}\right)^2 d\boldsymbol{x}\right]. \end{aligned} \tag{6}$$

The value within [] is non negative, and the sign of $S(t)$ is entirely determined by α.

3.3. Event Laplacian Product

We use "Event Laplacian Product" (ELP) as the focus detection function for the event-driven one-step AF, which is defined as:

$$\mathrm{ELP}(t) = -\sum \left(\nabla^2 I(t) \cdot E(t)\right), \tag{7}$$

where $I(t)$ denotes a grayscale image closest to the current time t, ∇^2 represents the Laplacian for $I(t)$, and $E(t)$ refers

to the event frame acquired at t. The definition of an event frame is as follows. Let the event stream from an event camera be denoted by $\{e_i\}$, where each event e_i is represented by the tuple (x_i, y_i, t_i, p_i). Here, (x_i, y_i) denotes the pixel location, t_i is the timestamp, and p_i represents the event polarity. The event frame $E(x, y, t - \Delta t, t)$, which accumulates events over the time interval $[t - \Delta t, t]$, is mathematically expressed as:

$$E(x, y, t - \Delta t, t) = \sum_i p_i \cdot \delta(x - x_i, y - y_i). \quad (8)$$

ELP(t) in Eq. (7) is the difference approximation of $S(t)$ in Eq. (6), where $E(t) \approx \frac{\partial}{\partial t} G(\boldsymbol{x}, t)$ and $I(t) \approx G(t)$. The gap between $G(t)$ and $I(t)$ is essentially whether h is the current defocus kernel. According to Eq. (6), $h(\boldsymbol{x}, t)$ can be replaced with $h(\boldsymbol{x}, t')$ at any given moment t' without affecting the sign of $S(t)$, as well as ELP(t). The gap between $E(t)$ and $\frac{\partial}{\partial t} G(\boldsymbol{x}, t)$ results in fluctuations in ELP values.

Figure 1 illustrates how the ELP varies with Δv as the system transitions from defocus to near-focus and back to defocus. As the focusing process approaches the focus position, the ELP value increases steadily. Upon reaching the focus position, the ELP undergoes a sharp "sign mutation" from positive to negative. Comparing event frames with the Laplacian of grayscale images at different Δv reveals the underlying causes of ELP value changes. As the system moves closer to the focus position, the coefficient α in Eq. (3) is negative, causing negative events to occur in regions where the Laplacian is positive, thereby resulting in a positive ELP value. When $|\Delta v|$ is large, the Laplacian values are generally low, and events are dispersed, resulting in a low ELP value. Approaching the focus position, Laplacian values increase, and the event frames sharpen progressively, leading to a surge in the absolute ELP value. Upon crossing the focus position, α in Eq. (3) turns positive, resulting in positive events occurring in regions of positive Laplacian, which induces an ELP "sign mutation".

The combination of the grayscale Laplacian and events gives ELP excellent properties as a focus detection function: (1) Its positive and negative values indicate whether the system is moving toward or away from the focus position, helping to avoid misguided "focus hunting"; (2) ELP is highly sensitive in identifying the focus position, with the abrupt "sign mutation" typically occurring within a single depth of focus; (3) ELP does not require **repeated searches** for the focus position, but only needs to **detect** the "sign mutation" **once**, making it a one-step AF solution.

3.4. ELP adaptive filter

In Eq. (8), the time interval Δt of event frames can be made very short (<1ms) to enhance the focus sensitivity. However, if Δt is too short, focus events may become susceptible to noise, leading to local fluctuations in the ELP. To address this, we introduce an adaptive ELP filter to suppress fluctuations. The filter is defined as follows:

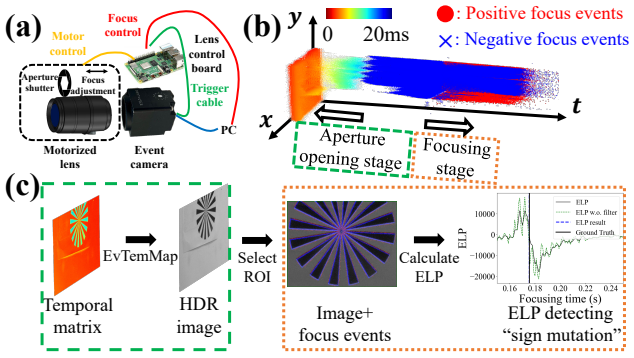


Figure 2. Event-driven one-step AF setup and pipeline. (a) Hardware setup. (b) Collected real events. (c) The pipeline of grayscale image acquisition and ELP value calculation.

1. Calculate the average of the past W collected ELP values, $\overline{ELP} = \frac{1}{W} \sum_{i=1}^{W} ELP_i$.
2. If $\left|ELP_{\text{now}} - \overline{ELP}\right| < ELP_{\text{thd}}$, return the filtered value $ELP_{\text{filtered}} = S \cdot ELP_{\text{now}} + (1 - S) \cdot \overline{ELP}$; otherwise, return the original value $ELP_{\text{filtered}} = ELP_{\text{now}}$.

The ELP adaptive filter determines whether to apply filtering based on the comparative judgment, which preserves the steepness of the ELP at the "sign mutation" while smoothing out local fluctuations elsewhere. The smoothing factor S, where $S \in [0, 1]$, controls the level of smoothness, with smaller values of S resulting in greater smoothness. The window size W defines the filter's visible range. The black solid line in Fig. 1 shows the filtered ELP, which removes local fluctuations in the green dashed line while preserving the steepness at the "sign mutation" point.

3.5. Event-only one-step AF pipeline

The event-only one-step AF pipeline consists of two stages: (1) **Aperture opening stage**: EvTemMap [2] is applied to capture a grayscale image; (2) **Focusing stage**: After selecting the focus region of interest (ROI), focus events are captured to compute ELP values and detect "sign mutation".

Figure 2 (a) shows the ELP hardware setup. Figure 2 (b) shows real events captured with a Prophesee EVK4 (an event-only camera), with 20 ms for the aperture opening and 100 ms for the focusing stage. The IPEs from aperture opening stage are color-coded by timestamp, while focus events are marked with blue 'x' for negative events and red 'o' for positive events. To illustrate the entire variation in ELP, we also capture focus events during the additional defocusing process. As shown in Figure 2 (c), in the aperture opening stage, following the principle of EvTemMap [2], we capture only the IPE from each pixel, forming a "Temporal matrix". After real-time EvTemMap, the "Temporal matrix" is converted into an HDR grayscale image, allow-

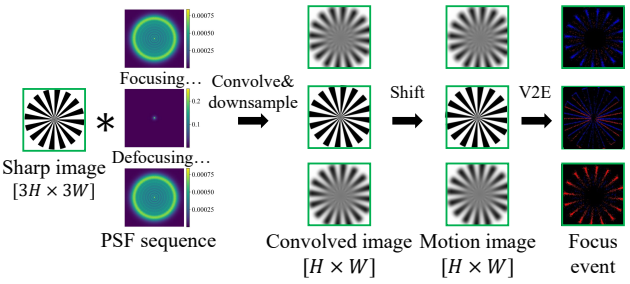


Figure 3. The pipeline of PSF-based focus event simulator.

ing the user to select the focus ROI. The focus motor then begins to move, and the event camera continuously captures focus events, which are combined with the previously obtained grayscale image to compute ELP values. During the focusing stage, the AF system continuously monitors changes in ELP: (1) If the ELP value is negative, the motor reverses direction toward the focus position; (2) When the ELP value shows an abrupt "sign mutation", indicating the focus position has been reached, the motor stops.

Notably, event cameras like the DAVIS346 can capture grayscale images directly and output them in sync with events, eliminating the need for an aperture opening stage.

4. Experiment

We compare the proposed ELP method with two event-driven AF methods EGS [10] and PBF [1] on synthetic datasets as well as the DAVIS346 and Prophesee EVK4 datasets, demonstrating the accuracy and efficiency of ELP.

4.1. Synthetic experiment setup

We develop a PSF-based focus event simulator to accurately generate focus events of real lenses, since the blur kernels of a real lens during focusing are determined by the PSFs at different Δv, which, due to lens aberrations, differ from ideal Gaussian blur kernels.

Simulator Overview. Figure 3 illustrates the pipeline for generating synthetic PSF-based focus events. The process begins by loading a sharp grayscale image with a resolution of $[3H \times 3W]$. This image is then convolved with $psf(\Delta v)$ at each defocus position Δv to generate the corresponding image, which is subsequently downsampled to a resolution of $[H \times W]$. In the synthetic dataset, H and W are equal to 200 pixels. To simulate motion during focusing, we translate the convolved image before generating events using the event simulator ("V2E") [7]. The focus event stack spans 1 second: ELP utilizes only the first 0.5 seconds of events (focusing process), while EGS and PBF use the entire 1-second focus event stack (focusing and defocusing process).

PSF characteristics. The synthetic dataset includes four groups of PSF sequences, corresponding to four fields of view (FoV), with each sequence containing 1001 PSFs sampled uniformly within the range of $\Delta v \in [-400, 400]$ μm. The RMS radius of the PSF at maximum defocus is 10 times that at the focus position, indicating a challenging initial defocus. For further details on the PSF, please refer to the supplementary material.

Motion simulation. The motion vector, $\mathbf{t}$, for each frame consists of two components: (1) random jitter v_{jitter} and (2) constant motion v_{motion}, expressed as $\mathbf{t} \sim \mathcal{N}(v_{\text{motion}}, v_{\text{jitter}}^2)$. In the synthetic dataset, we provide two types of motion parameters: moderate and violent. For the moderate motion, $v_{\text{motion}} = [3, 3]$ pixels/s and $v_{\text{jitter}} = [20, 20]$ pixels/s, whereas for the violent motion, $v_{\text{motion}} = [3, 3]$ pixels/s and $v_{\text{jitter}} = [100, 100]$ pixels/s.

Brightness simulation. Brightness affects the frame rate of grayscale image capture, impacting the acquisition of the Laplacian in the ELP method. Under normal brightness (>100 lux), the DAVIS346 camera captures images at up to 50 frames per second (FPS). However, at lower brightness (<1 lux), the grayscale image frame rate drops to 20 FPS or lower. We evaluate the performance of ELP at both 50 FPS and 20 FPS. Additionally, since the event-only one-step AF system mentioned in Sec. 3.5 uses only a single defocus image, we also simulate this scenario.

4.2. Synthetic experiment result

Method	Motion state		
	Static	Moderate	Violent
EGS	33.31	29.33	21.78
PBF	4.93	3.99	2.69
ELP(1FPS)	**2.00**	3.66	3.66
ELP(20FPS)	2.00	2.40	2.97
ELP(50FPS)	2.00	**1.51**	**2.26**

Table 1. MAE comparison on synthetic datasets (in μm). ELP method is tested with various grayscale frame rates under three lighting conditions. The best performances are marked in **bold**.

In the synthetic dataset of 84 cases, the EGS method yields the highest mean absolute error (MAE), as shown in Tab. 1. In contrast, both the PBF and ELP methods achieve near-optimal results across various motion states. Detailed results for each case in the synthetic dataset are provided in the supplementary material, where the focusing errors for both PBF and ELP remain within one depth of focus (16 μm) for each case. However, the EGS method fails to provide a focus position in 28.6% of cases, and in 68.3% of the remaining cases, its error exceeds one depth of focus. With its one-step AF principle, ELP eliminates the need for defocusing and returning to the target, reducing focusing time by two-thirds compared to PBF and EGS.

Figure 4 presents a visual comparison of three event-driven AF methods in a violent motion scenario, where ELP utilizes only one grayscale frame (1 FPS) for Laplacian information. In this case, both PBF and ELP achieve sharp

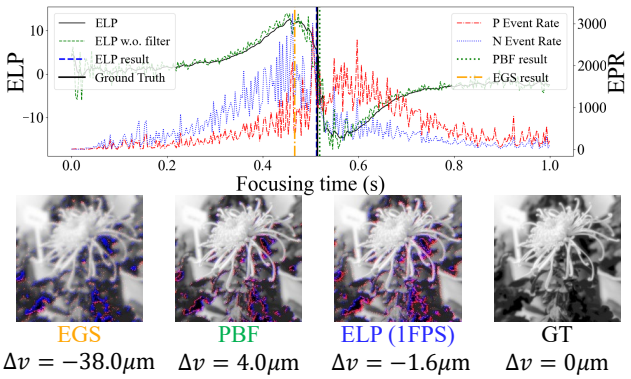


Figure 4. Visualization of a violent shake scene. Top row: temporal variation of ELP and EPR across the entire focus stack. Bottom row: grayscale images at different Δv overlaid with focus events.

focus, while the blur kernel of the EGS method exceeds two pixels, as shown in the bottom row. The top row reveals that the severe jitter causes significant localized fluctuations in both the EPR and the ELP curves. The ELP adaptive filter effectively suppresses these fluctuations while preserving the steepness of the "sign mutation" point, ensuring the accuracy of ELP. In instances of random violent jitter, EGS incorrectly identifies the location with the highest ER as the focus position, whereas PBF accurately locates the center of symmetry of the EPR, closely aligning with the Ground Truth (GT) focus position.

4.3. Real experiment setup

Focusing hardware and configuration. In real experiments, a motorized focusing lens is used for the focusing process, with DAVIS346 and Prophesee EVK4 cameras capturing events (and, if available, grayscale images). The focus ring of the motorized lens is driven by a stepper motor. When the lens reaches the GT focus position, a trigger signal is sent to the AF control unit to mark the GT timestamp during focusing. The GT focus position is determined by locating the point where the blinking focus star appears sharpest. The complete focus event stack—from defocus to near-focus and back to defocus—spans approximately 200 ms, starting from a significant Δv of 200 μm.

DAVIS dataset. In the DAVIS dataset, time-synchronized event and grayscale images are captured simultaneously. The maximum frame rate for grayscale images is 50 FPS, but it can drop to as low as 20 FPS under low-light conditions. ELP uses real-time events and the latest grayscale frame to compute the value and detects in real time whether the "sign mutation" has happened. In contrast, EGS and PBF require capturing the entire focus event stack, performing an algorithmic search for the focus position, and then driving the motor to that position.

EVK4 dataset. Since EVK4 only provides events, we build a **event-only one-step AF system** (*cf*. Sec. 3.5). ELP computes the focus detection function using a single frame of the defocus image acquired from the aperture opening stage, along with the subsequent real-time focus events. In contrast, EGS and PBF do not use events from the aperture opening stage; instead, they rely on the event stack from the entire focusing stage. The aperture is opened by the motorized lens's high-speed motor, transitioning from fully closed to fully open within 20 ms.

4.4. Real experiment result

One-step AF visualization. The supplementary videos provide a visualization of the one-step AF process of the ELP method on real datasets, showcasing its ability to drive directly to the focus position with precision and efficiency.

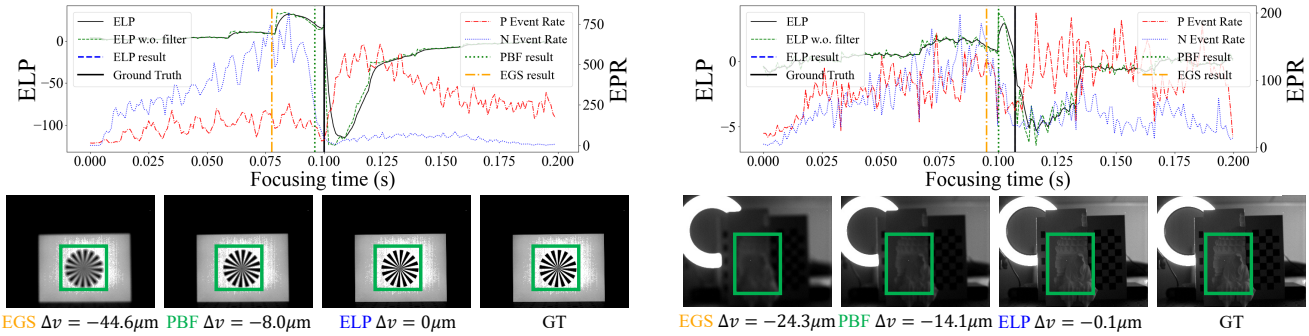
Quantitative results. Table 2 details the quantitative focusing errors of the EGS, PBF, and our ELP methods on the real datasets. The motorized focusing len is considered accurately focused if the focusing error is within 6 μm (*i.e*. one depth of focus). The ELP method achieves accurate focus in every case for both the DAVIS and EVK4 datasets, while the PBF method achieves accurate focus in approximately 43% of the cases, and the EGS method fails to achieve accurate focus in any case. Table 3 summarizes the MAE of the three event-driven AF methods on both datasets, showing that the focusing error of the ELP method is reduced by 24 times compared to the state-of-the-art PBF method on the DAVIS346 dataset and by 22 times on the EVK4 dataset.

Speed analysis. The focusing speed analysis for the three event-driven AF methods is summarized in Tab. 3, where "Runtime" refers to the algorithm's execution time, and "Focusing time" includes the entire focusing process, encompassing motor operation and data acquisition. In terms of algorithm runtime, the PBF algorithm performs best on both datasets, with a complexity of $O\left(\frac{1}{\Delta t}\right)$ [1], where $1/\Delta t$ is the sampling rate of event frames. The EGS algorithm performs well on the DAVIS dataset but worst on the EVK4 dataset, due to its complexity of $O(N_e)$ [10], where N_e denotes the ER. Our ELP algorithm demonstrates moderate runtime performance on both datasets, with a complexity of $O\left(\frac{k}{\Delta t}\right)$, where k represents the number of non-zero pixels in the event frame. For the total time required for the complete focusing process, however, our ELP method is nearly three times faster than the other two methods. This efficiency results from ELP's one-step AF principle, which continuously detects in real time if the focus position is reached as the focus motor moves, rather than waiting for the full focusing trip to complete and then analyzing the entire focus stack before positioning the motor.

Impact of contrast, brightness, and motion. A high-contrast, bright, and static scene is a straightforward case for the focusing task, as seen in Scene 1 of Fig. 5. In this scene, ELP achieves precise focus, whereas PBF exhibits

dataset	scene	bright_dynamic			bright_static			dark_dynamic			dark_static		
		EGS	PBF	ELP	EGS	PBF	ELP	EGS	PBF	ELP	EGS	PBF	ELP
DAVIS	box	-10.8	-1.4	**0.3**	-33.1	-9.2	**0.2**	-27.1	**0.1**	0.5	-25.4	-7.2	**0.2**
	focus star	-26.7	2.4	**0.4**	-44.6	-8.0	**0.0**	-28.9	-3.0	**0.0**	-27.2	-3.4	**0.7**
	forest	-12.1	-3.1	**0.6**	-14.8	-6.4	**-0.7**	-11.5	1.4	**0.2**	-13.1	-5.9	**0.2**
	ghost	-21.2	-19.4	**0.6**	-54.1	-25.6	**0.4**	-24.3	-14.1	**-0.1**	-24.9	-7.7	**0.3**
	lens	-44.5	4.3	**0.7**	-42.3	3.4	**-0.2**	-39.1	-35.5	**0.2**	-37.8	6.7	**0.0**
	mountain	-12.4	1.1	**0.1**	-28.4	-7.2	**-0.4**	-13.6	2.1	**0.2**	-22.3	-5.0	**0.0**
	statue	-65.4	9.1	**0.1**	-38.7	-16.8	**0.2**	-26.4	-13.2	**0.2**	-25.1	-1.5	**0.6**
EVK4	box	-17.4	10.6	**0.6**	-32.4	-1.9	**0.1**	-15.6	13.7	**-0.3**	-26.2	4.7	**-1.3**
	focus star	29.8	-6.6	**0.4**	-26.0	4.0	**0.0**	86.9	5.9	**-0.1**	88.1	10.1	**0.1**
	forest	-16.3	-3.7	**0.3**	-9.1	6.1	**0.1**	-9.4	9.7	**-0.3**	-10.8	6.8	**-0.2**
	ghost	-38.4	18.4	**0.4**	-37.5	**0.1**	**0.1**	-6.2	35.3	**-0.7**	-16.5	7.2	**3.2**
	lens	-26.4	-4.3	**0.7**	126.3	-5.8	**0.2**	120.5	19.3	**1.3**	164.8	72.7	**-0.3**
	mountain	-20.2	-2.3	**-0.3**	-18.4	2.0	**0.0**	-15.4	7.7	**-0.3**	-44.2	40.4	**0.4**
	statue	-36.3	10.7	**0.7**	-37.3	7.0	**0.0**	-48.4	-7.3	**1.7**	-26.5	26.8	**1.8**

Table 2. Quantitative comparison of focusing errors across event-driven AF methods on each case of real datasets, measured in μm. The best performances are marked in **bold**.



Scene 1: Focus star (high-contrast), bright, static

Scene 2: Ghost (low-contrast), dark, dynamic

Figure 5. Visualization of two scenes from the DAVIS dataset. Top row: temporal variation of ELP and EPR across the entire focus stack. Bottom row: grayscale images corresponding to the focus positions determined by the three event-driven methods, alongside the GT. Green boxes indicate the focus ROI.

Dataset	Method	MAE (μm)	Runtime (ms)	Focusing time (ms)
DAVIS	EGS	28.42	10.81	310.81
	PBF	7.02	**2.37**	302.37
	ELP	**0.29**	12.14	**103.36**
EVK4	EGS	41.12	104.25	404.25
	PBF	12.54	**2.52**	302.52
	ELP	**0.57**	83.00	**112.97**

Table 3. Quantitative comparison of focusing performance. The best performances are marked in **bold**.

a focusing error exceeding one depth of focus, and EGS shows an even larger error, exceeding seven depths of focus. Observing the changes in ELP and EPR over time, the "sign mutation" point of the ELP curve is very close to the symmetric position of the positive and negative EPR curves. Recalling Eq. (8), ELP, like PBF, also incorporates event polarity information. The **search** for the symmetry center of EPR in PBF is transformed into the **detection** of the "sign mutation" in ELP using the additional spatial texture information provided by the Laplacian, enabling one-step AF. In the absence of optical flow, the principle of EGS cannot hold, which explains its failure.

In a low-contrast, dark, and dynamic scene (Scene 2 in Fig. 5), the signal-to-noise ratio of the focus event drops sharply, resulting in significant fluctuations in both ELP and EPR curves. Under these challenging conditions, EGS demonstrates a focusing error of four depths of focus, whereas PBF exceeds two depths of focus, resulting in noticeable blurring. In contrast, ELP consistently achieves precise focus. This example highlights how incorporating Laplacian information from grayscale images enhances the event-driven AF robustness in extreme scenarios.

Event-only one-step AF. Figure 6 shows two scenes from the EVK4 dataset captured by the event-only one-step AF system. Although the event-only system is limited to acquiring a single defocus grayscale frame, it benefits from EvTemMap's high dynamic range, ultra-high grayscale resolution, and large depth of field [2]. The increased dynamic range allows ELP to utilize more texture information, while

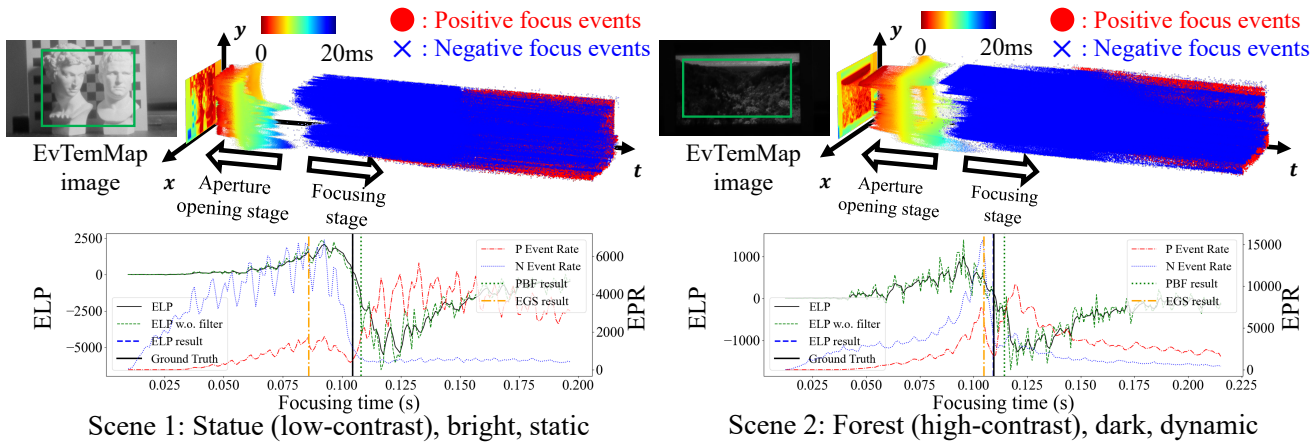


Figure 6. Visualization of two scenes from the EVK4 dataset. Top row: visualization of raw event data and initial defocus grayscale images obtained from EvTemMap [2], with green boxes indicating the focus ROI. Bottom row: temporal variation of ELP and EPR.

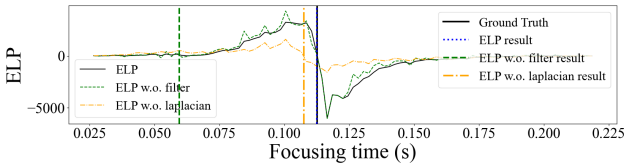


Figure 7. Visualization of ELP curves from the ablation study.

the enhanced grayscale resolution improves the accuracy of Laplacian computations. Additionally, the larger depth of field enables the frame to serve as a sharp texture reference for ROIs at different focus positions, similar to an all-in-focus frame. The objects captured in Scene 1 of Fig. 6 and Scene 2 of Fig. 5 are both low-contrast plaster statues. However, the high grayscale resolution and dynamic range of EvTemMap provide a more accurate Laplacian, allowing ELP to perform effectively even with only a single frame for reference. In Scene 2 of Fig. 6, characterized by richer textures and more motion events, the ELP curve shows greater local fluctuations. The ELP adaptive filter effectively reduces most of these fluctuations, ensuring robust results. Comparing the two scenes in Fig. 6, the richer texture and stabilized optical flow in Scene 2 enable EGS to achieve better focusing results. Both PBF and ELP, which incorporate event polarity, consistently achieve accurate focus across different scenarios.

Dataset	ELP	w.o. filter	w.o. Laplacian
DAVIS	0.29	2.12	2.48
EVK4	0.57	19.85	7.81

Table 4. Ablation experiment results. Metric: MAE (μm).

Ablation study. Table 4 summarizes the results of ablation experiments on real datasets, while Fig. 7 illustrates the ELP curves for a specific case (mountain, bright, dynamic) in the EVK4 dataset under two ablation settings. In the ELP method, the adaptive filter is crucial for robustness, especially in the EVK4 dataset, which uses a single grayscale frame. As shown by the green dashed line in Fig. 7, removing this filter introduces noise fluctuations, causing premature false focus detections. Replacing the Laplacian with a uniform 1 matrix flattens the orange dashed-dotted ELP curve, leading to an earlier "sign mutation". Without Laplacian information, ELP degenerates into the difference between negative and positive event rates, losing the ability to determine the focus direction based on sign. Consequently, in 40% of cases in the DAVIS dataset, the ELP without Laplacian provides incorrect focus adjustment direction, resulting in "focus hunting". Additionally, for the event-only EVK4 dataset, we ablate the EvTemMap method and instead use E2VID to obtain grayscale Laplacian information. Detailed results are provided in the supplementary material.

5. Discussion

We introduce the first event-driven, one-step AF method, the Event Laplacian Product (ELP), which reduces focus time to one-third of existing event-driven methods and resolves the issue of "focus hunting". Experiments on synthetic data and two real-world event camera datasets across diverse lighting and motion conditions demonstrate that ELP consistently achieves precise focus within one depth of field. Compared to the state-of-the-art event-driven method PBF, ELP reduces focusing error by 24 times on the DAVIS346 dataset and by 22 times on the EVK4 dataset.

ELP is typically fast and accurate across most scenarios but faces challenges under extreme high-speed motion in event-only settings. Future work could focus on adaptive "sign mutation" detection to enhance ELP's robustness, ensuring reliable performance in demanding conditions.

Acknowledgment

This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grant No. LZ24F050003, in part by the National Key R&D Program of China, under Grant No. 2022YFF0705500 and No. 2022YFB3206000.

References

[1] Yuhan Bao, Lei Sun, Yuqin Ma, Diyang Gu, and Kaiwei Wang. Improving fast auto-focus with event polarity. *Optics Express*, 31(15):24025–24044, 2023. 1, 2, 3, 5, 6

[2] Yuhan Bao, Lei Sun, Yuqin Ma, and Kaiwei Wang. Temporal-mapping photography for event cameras. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024. 2, 3, 4, 7, 8

[3] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[4] Zhou Ge, Haoyu Wei, Feng Xu, Yizhao Gao, Zhiqin Chu, Hayden K-H So, and Edmund Y Lam. Millisecond autofocusing microscopy using neuromorphic event sensing. *Optics and Lasers in Engineering*, 160:107247, 2023. 2

[5] Chenzi Guo, Zelong Ma, Xu Guo, Wenxian Li, Xinda Qi, and Qinglei Zhao. 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. 2

[6] 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

[7] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1312–1321, 2021. 5

[8] Dongyao Jia, Chuanwang Zhang, Nengkai Wu, Jialin Zhou, and Zhigang Guo. Autofocus algorithm using optimized laplace evaluation function and enhanced mountain climbing search algorithm. *Multimedia Tools and Applications*, 81(7):10299–10311, 2022. 2

[9] Eric Krotkov. Focusing. *International Journal of Computer Vision*, 1(3):223–237, 1988. 2

[10] Shijie Lin, Yinqiang Zhang, Lei Yu, Bin Zhou, Xiaowei Luo, and Jia Pan. Autofocus for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16344–16353, 2022. 1, 2, 5, 6

[11] Matthew Lofroth and Ebubekir Avci. Auto-focusing approach on multiple micro objects using the prewitt operator. *International Journal of Intelligent Robotics and Applications*, 2(4):413–424, 2018. 2

[12] Hanyue Lou, Minggui Teng, Yixin Yang, and Boxin Shi. All-in-focus imaging from event focal stack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17366–17375, 2023. 2

[13] Chunhong Mo and Bo Liu. An auto-focus algorithm based on maximum gradient and threshold. In *2012 5th International Congress on Image and Signal Processing*, pages 1191–1194. IEEE, 2012. 2

[14] Henry Pinkard, Zachary Phillips, Arman Babakhani, Daniel A Fletcher, and Laura Waller. Deep learning for single-shot autofocus microscopy. *Optica*, 6(6):794–797, 2019. 2

[15] Nicholas Owen Ralph, Darren Maybour, Alexandre Marcireau, Nik Dennler, Sami Arja, Nimrod Kruger, and Gregory Cohen. Active neuromorphic space imaging and focusing using liquid lenses. *Authorea Preprints*, 2024. 2

[16] Przemysław Śliwiński and Paweł Wachel. A simple model for on-sensor phase-detection autofocusing algorithm. *Journal of Computer and Communications*, 1(06):11, 2013. 2

[17] Minggui Teng, Hanyue Lou, Yixin Yang, Tiejun Huang, and Boxin Shi. Hybrid all-in-focus imaging from neuromorphic focal stack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[18] Fengting Yang, Xiaolei Huang, and Zihan Zhou. Deep depth from focus with differential focus volume. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12642–12651, 2022. 2

[19] Yupeng Zhang, Liyan Liu, Weitao Gong, Haihua Yu, Wei Wang, Chongying Zhao, Peng Wang, and Toshitsugu Ueda. Autofocus system and evaluation methodologies: A literature review. *Sensors & Materials*, 30, 2018. 1