

One-Step Event-Driven High-Speed Autofocus

Supplementary Material

6. PSF-based focus event simulator details

Figure 8 shows the details of the real 50mm F2.0 lens used in the synthetic dataset, where the top row shows the structure of the lens. Of interest is the PSF of this lens at different Δv , as shown in the bottom row. At the focus position, the RMS radius of the PSF is less than 1, *i.e.*, the vast majority of the energy is concentrated in the 3×3 region within the white dashed box. According to the paragraph **Simulator Overview** in Sec. 4.1, the convolved image is 1/3 downsampled, so the PSF at the focus position can be considered an ideal Dirac function $\delta(x)$.

As shown by the PSFs for $\Delta v = -400\mu\text{m}$ and $\Delta v = 400\mu\text{m}$ for each FoV in Fig. 8, the PSFs of the real lens differ from the ideal Gaussian blur kernel. Additionally, for a real lens, the size of the blur kernel (as measured by the RMS radius of the PSF) may not necessarily change linearly with Δv , as depicted in the top row of Fig. 9.

7. Depth of focus

The concept of “depth of focus” has been referenced multiple times in the main text. Precise focusing is defined as a focusing error within one depth of focus. For example, considering the lens used for the synthetic dataset with a depth of focus of $16\mu\text{m}$, the blur kernel can be approximated as a Dirac function, $\delta(x)$, whenever the focusing error falls within this range. As illustrated in the bottom row of Fig. 9, the images within the yellow box are indistinguishable from the sharp original image, indicating precise focusing.

However, aiming for a smaller focus error within the depth of focus offers clear benefits: it brings the lens closer to the center depth within the focus ROI, allowing objects just in front of and behind the main focus point to stay sharp. For instance, when focusing on a face, it’s preferable for features like the nose tip and ears to be as clear as the eyes, rather than achieving sharpness for the eyes and nose while letting the ears become blurred.

8. Event-only one-step AF details

In the event-only one-step AF system, Event-driven Temporal-mapping Photography (EvTemMap) [2] plays a crucial role by providing ELP with a single grayscale reference frame for accurate Laplacian computation. As discussed in Sec. 4.4, EvTemMap images exhibit a high dynamic range, ultra-high grayscale resolution, and an extended depth of field, all of which facilitate precise Laplacian acquisition in ELP. The high dynamic range and grayscale resolution arise from temporal mapping, where

each microsecond timestamp is mapped to one grayscale level, resulting in nearly 20,000 levels over a 20 ms exposure. The extended depth of field is achieved through a specialized transmittance modulation approach: in our setup, an aperture shutter opens progressively from fully closed. In this setup, brighter areas correspond to smaller apertures, which, in turn, produce a greater depth of field.

Comparing Scene 2 in Fig. 5 with Scene 1 in Fig. 6, the dynamic range of the grayscale image captured by the DAVIS346 APS sensor is notably lower than that of the EvTemMap image, which reveals detailed texture. The higher grayscale resolution in EvTemMap provides more refined Laplacian information for ELP, resulting in generally higher ELP values for the EVK4 dataset compared to the DAVIS dataset. Additionally, the extended depth of field in EvTemMap makes the defocus image appear sharper, resembling an all-in-focus image. This enhanced sharpness provides more precise texture information, thereby increasing the steepness of the ELP “sign mutation” at the focus position. Although the grayscale image has a large depth of field, the focus events still correspond to a shallow depth of field, ensuring accurate focus position detection. In a darker environment, such as Scene 2 in Fig. 6, the defocus image from EvTemMap appears less sharp. Consequently, the ELP curve shows a slightly reduced steepness at the “sign mutation” point, potentially increasing focusing error. However, as in most focusing scenarios, beginning with a small defocus amount enables the event-only one-step autofocus system to function effectively, even under low-light conditions.

9. Detailed results on the synthetic dataset

Table 5 details the focusing errors of the EGS, PBF, and ELP methods for the 84 scenes of the synthe dataset. In the synthetic dataset, the ground truth for the focus position is the point with the smallest PSF RMS radius. With a focus depth of $16\mu\text{m}$, the focus is considered accurate if the focus error remains within $16\mu\text{m}$ and the blur kernel radius is smaller than one pixel. As shown in Tab. 5, all three ELP setups produce accurate focus results, as does the PBF. In contrast, the EGS achieves accurate focus in only 17.8% of cases, and in 28.6% of cases, it fails to provide any valid focus results (denoted by ‘/’).

10. Ablation study on reconstruction quality.

The quality of image reconstruction greatly impacts ELP. We compare two approaches: Motion E2VID, which applies E2VID to the ego-motion events collected before the

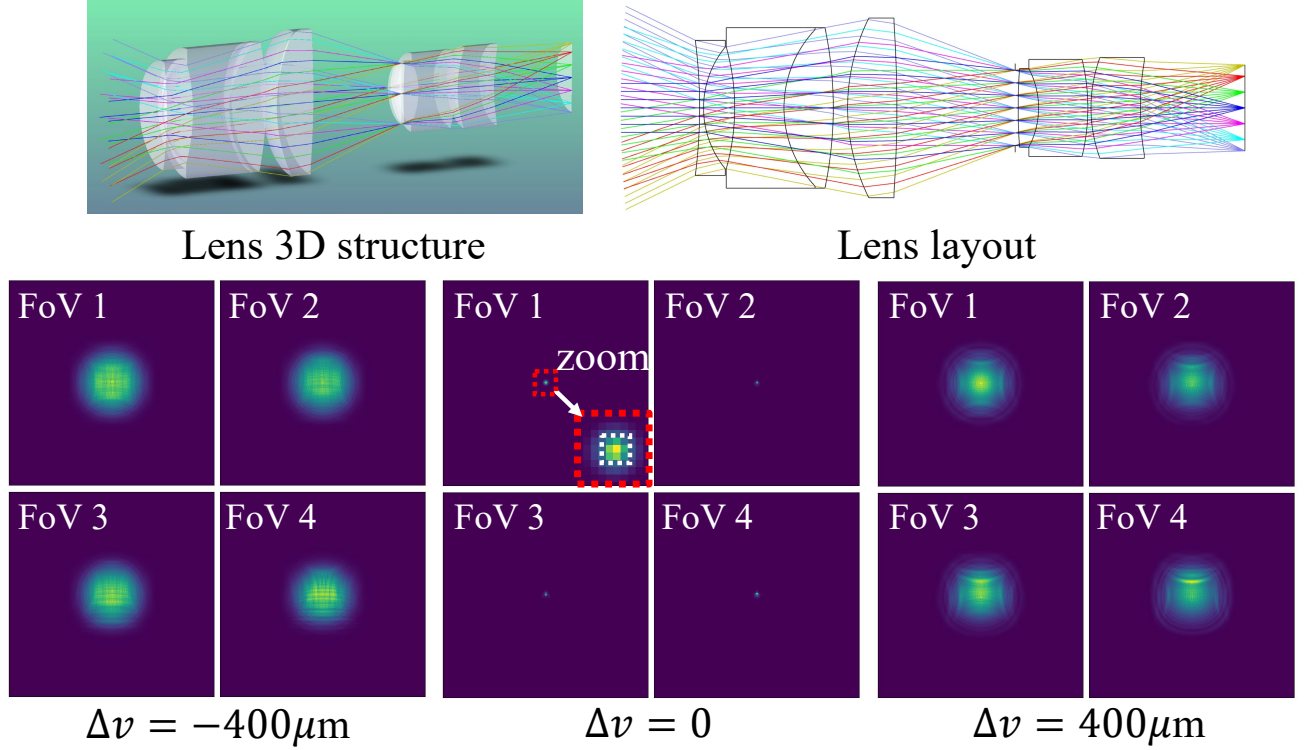


Figure 8. Details of the real 50mm F2.0 lens used for the synthetic dataset. Top row: lens structure. Bottom row: PSFs of the 4 FoVs at different Δv . At the focus position position, the RMS radius of the PSF is less than 1. Since the convolved images are 1/3 downsampled, the blur kernel is an ideal Dirac function $\delta(\mathbf{x})$ at the focus position.

focus stage, and Focus E2VID, which applies E2VID to focus events of the entire stack. Frames with the lowest NIQE are selected as reference frames for ELP. As shown in Fig. 10, Motion E2VID fails due to degraded texture, while Focus E2VID, though improved, introduces grayscale errors, increasing MAE on the EVK4 dataset from $0.57\mu\text{m}$ to $43.2\mu\text{m}$. In contrast, EvTemMap not only provides higher-quality reference frames, but also offers a more streamlined workflow: Opening a closed aperture before capturing a snapshot aligns better with user habits than inducing ego-motion or pre-capturing a full focus stack.

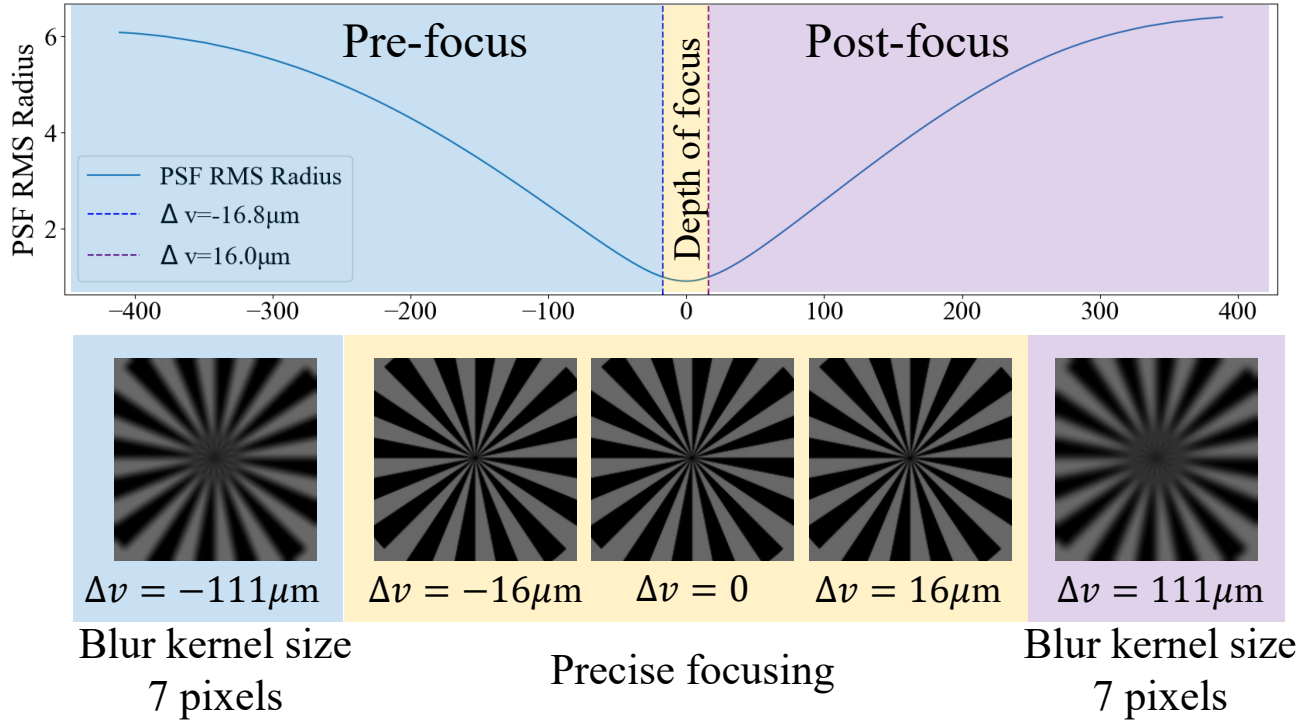
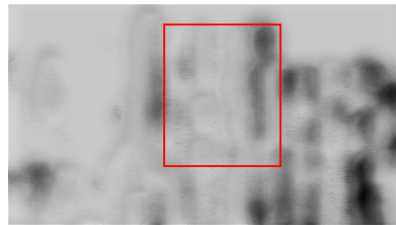
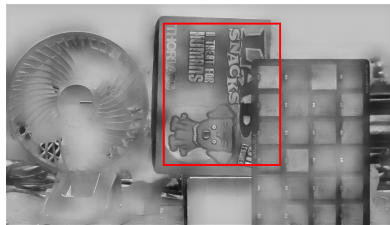


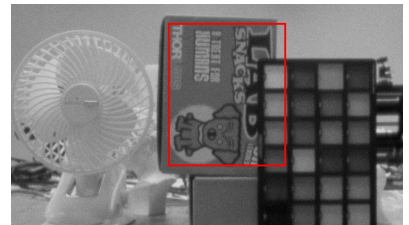
Figure 9. Top row: PSF RMS radius changes with Δv . Bottom row: the visual explanation of the depth of focus. The yellow box represents the depth of focus region, while the blue and purple boxes indicate the pre-focus and post-focus regions, respectively.



Motion E2VID:
No focus found



Focus E2VID:
-9.0μm



EvTemMap:
0.7μm

Figure 10. Ablation study on reconstruction quality.

Scene	FoV	EGS			PBF			ELP(1FPS)			ELP(20FPS)			ELP(50FPS)		
		S	M	V	S	M	V	S	M	V	S	M	V	S	M	V
focus-board	1	/	/	/	1.20	1.60	0.80	-8.00	-8.00	-6.40	-2.40	-5.60	-0.80	-1.60	-3.20	-4.80
	2	-44.80	/	-57.60	2.40	4.00	0.40	-3.20	-12.80	-8.00	-0.80	-0.80	-2.40	0.00	-1.60	-1.60
	3	-45.60	-40.00	-33.87	3.60	3.60	3.20	-4.80	-6.40	-6.40	-2.40	-0.80	-0.80	-2.40	0.00	0.00
	4	-36.00	-50.80	/	10.00	10.00	7.20	3.20	1.60	6.40	-0.80	0.80	2.40	0.00	0.00	1.60
dove	1	-31.20	/	/	-2.40	-2.80	-3.20	-2.40	2.40	-0.80	-1.60	-1.60	-6.40	-0.80	-2.40	-2.40
	2	/	-12.80	-6.00	5.60	0.80	-2.40	0.00	-1.60	-3.20	2.40	0.80	-0.80	3.20	1.60	1.60
	3	/	/	/	4.80	4.00	4.00	1.60	0.00	0.00	5.60	4.00	7.20	6.40	4.80	4.80
	4	/	/	-36.80	1.60	4.80	1.60	0.80	-4.00	-5.60	-3.20	-3.20	-4.80	-0.80	-0.80	-0.80
cat	1	-38.40	-16.80	/	-3.20	-3.20	-4.00	0.80	-0.80	-4.00	-1.60	-4.80	-3.20	-2.40	-2.40	-4.00
	2	/	/	/	6.40	0.80	-4.00	1.60	-3.20	-1.60	0.80	0.80	0.80	1.60	1.60	0.00
	3	-31.20	/	-31.20	5.60	4.00	4.00	3.20	1.60	0.00	4.00	2.40	0.80	4.80	3.20	1.60
	4	/	/	-40.80	4.80	8.00	4.80	2.40	0.80	2.40	0.00	-1.60	-4.80	-0.80	-0.80	-0.80
flower	1	-38.80	-49.20	2.20	4.00	2.40	0.80	0.00	-3.20	-1.60	0.80	-2.40	-2.40	1.60	0.00	-1.60
	2	-38.40	-43.60	-33.76	6.40	4.00	-4.80	0.00	-4.80	1.60	0.80	-2.40	0.80	0.00	-1.60	1.60
	3	-50.40	-50.00	-25.92	4.80	4.00	2.40	0.00	-1.60	-4.80	2.40	0.80	-0.80	3.20	1.60	-0.80
	4	/	-44.40	-38.00	10.40	10.40	4.00	4.80	0.00	-1.60	0.80	-0.80	0.80	1.60	0.00	0.00
leaf	1	-38.80	-15.60	8.64	4.00	0.80	-0.80	0.00	3.20	1.60	0.80	-4.00	-4.00	0.00	-1.60	-3.20
	2	-36.80	-34.40	29.87	7.20	-2.40	-4.00	0.00	-8.00	-9.60	0.80	-2.40	0.80	1.60	0.00	4.80
	3	-40.80	/	-32.40	7.20	4.00	-5.60	0.00	-1.60	-6.40	4.00	0.80	0.80	6.40	3.20	4.80
	4	-29.60	-29.60	14.20	11.20	13.60	2.40	4.80	0.00	9.60	-0.80	-0.80	7.20	1.60	0.00	0.00
chair	1	-4.27	-0.16	7.73	0.40	0.80	1.20	1.60	-1.60	0.00	-0.80	-5.60	-8.80	-1.60	-3.20	-3.20
	2	-5.80	-5.71	7.70	-0.80	-0.80	0.00	-1.60	-4.80	-4.80	-2.40	-4.00	-4.00	-3.20	-3.20	-3.20
	3	-29.76	-34.40	-7.36	0.80	1.60	0.80	-3.20	-6.40	-3.20	-2.40	-0.80	-4.00	-0.80	-0.80	0.00
	4	-3.20	-3.60	-14.40	4.00	4.00	1.20	0.00	-3.20	6.40	-2.40	-4.00	-4.00	-1.60	-1.60	-3.20
grass	1	/	-21.20	15.60	4.00	0.00	0.80	3.20	-3.20	-1.60	2.40	-4.00	-0.80	1.60	-1.60	-1.60
	2	-37.60	-32.00	-1.60	7.20	-5.60	-4.00	1.60	-8.00	1.60	4.00	-2.40	5.60	3.20	0.00	-1.60
	3	-32.00	-30.67	11.73	3.60	-2.40	-2.00	-1.60	-4.80	-1.60	2.40	-0.80	-0.80	3.20	0.00	4.80
	4	52.80	-42.40	/	10.40	7.20	0.80	1.60	-4.80	1.60	-2.40	-4.00	-2.40	0.00	-1.60	4.80

Table 5. Detailed results of the synthetic dataset, measured in μm . S for Static, M for Moderate motion, V for Violent motion. ‘/’ means that focus position fails to be identified.