CodedEvents: Optimal Point-Spread-Function Engineering for 3D-Tracking with Event Cameras

Sachin Shah Matthew A. Chan Haoming Cai Jingxi Chen Sakshum Kulshrestha
Chahat Deep Singh Yiannis Aloimonos Christopher A. Metzler

University of Maryland, College Park
shah2022@umd.edu

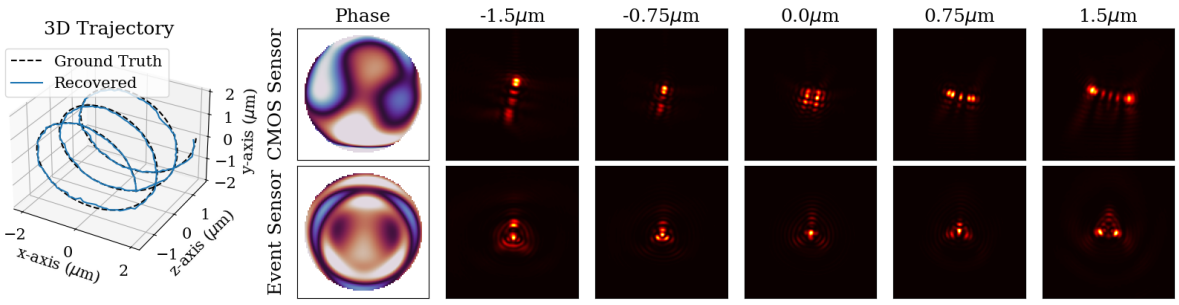


Figure 1. **CodedEvent Tracking.** Left: example recovered trajectory using designed optics for an event camera. Right: top row, optimal phase mask design and PSFs for a CMOS sensor, bottom row, our optimal phase mask design and PSFs for an event sensor.

Abstract

Point-spread-function (PSF) engineering is a well-established computational imaging technique that uses phase masks and other optical elements to embed extra information (e.g., depth) into the images captured by conventional CMOS image sensors. To date, however, PSF-engineering has not been applied to neuromorphic event cameras; a powerful new image sensing technology that responds to changes in the log-intensity of light.

This paper establishes theoretical limits (Cramér Rao bounds) on 3D point localization and tracking with PSF-engineered event cameras. Using these bounds, we first demonstrate that existing Fisher phase masks are already near-optimal for localizing static flashing point sources (e.g., blinking fluorescent molecules). We then demonstrate that existing designs are sub-optimal for tracking moving point sources and proceed to use our theory to design optimal phase masks and binary amplitude masks for this task. To overcome the non-convexity of the design problem, we leverage novel implicit neural representation based parameterizations of the phase and amplitude masks. We demonstrate the efficacy of our designs through extensive simulations. We also validate our method with a simple prototype.

1. Introduction

Single-molecule localization microscopy (SMLM) is a vital tool for resolving nano-scale structures with applications in analysis of protein clusters [39], cell dynamics [62], and electromagnetic effects [30]. Traditional SMLM experiments are limited by the slow capturing process of frame-based CMOS sensors, preventing use in capturing high-speed, dynamic interactions. Recently, [9] showed event cameras are key to enabling high-speed 2D SMLM.

In contrast to traditional CMOS cameras, event cameras are an emerging class of bio-inspired neuromorphic sensors that operate with a high temporal resolution on the order of μs. These sensors are comprised of an asynchronous pixel array, where each pixel records an event when the log intensity change exceeds a set threshold. In addition to having kilohertz time resolution, these sensors are low-power, resistant to constant background noise, and can operate over a high dynamic range [14]. Already, these sensors have proven useful in a range of applications including object tracking [4, 60], gesture recognition [3, 32], and robotics [24, 48].

Just as PSF-engineering allows one to extract additional information using conventional CMOS sensors [52], we believe that event-camera-specific PSF engineering will be the key to enabling high-speed 3D SMLM with event cameras.

Unfortunately, existing PSF design theory is not equipped for the event space. In this work, we bridge this gap by developing Cramér Rao Bounds on 3D position estimation for event camera measurements. Leveraging these bounds, we subsequently develop a novel implicit neural representation for optical elements to design components with improved 3D particle localization capabilities.

Specifically, our principal contributions are as follows:

- We derive the Fisher Information and Cramér Rao Bounds for event camera measurements parameterized by 3D spatial positions.
- We develop novel implicit neural representations for learning both amplitude and phase masks.
- We identify new phase and amplitude designs for optimally encoding 3D information with event cameras.
- We demonstrate in simulation that our designs outperform existing methods at 3D particle tracking.

2. Related Work

2.1. Coded Optics

Specialized lenses have been shown to encode additional depth information in CMOS image frames. A 'coded aperture' can produce depth-dependent blurs that enable one to extract depth by looking at the per-pixel defocus pattern [33]. Future works extend the 'depth from defocus' idea by leveraging information theory to design an optimal lens [27, 52]. More recently, researchers have proposed optimizing optical parameters in conjunction with a neural network reconstruction algorithm in an 'end-to-end' fashion. This joint-optimization problem is difficult to optimize due to local minima. Many works have discussed mask parameterizations to stabilize optimization: Zernike basis [10, 64] and rotationally symmetric [25]. However, direct pixel-wise methods should be preferred due to their expressiveness [36]. Dynamic pixel-wise masks have been proposed as a training stabilization mechanism [50]. Specialized optics have been explored for other applications such as super resolution [55], high-dynamic-range imaging [40], hyperspectral sensing [34], and privacy-preservation [22]. To our knowledge, PSF engineering specifically for event-based sensors has been relatively unexplored.

2.2. Microscopy Tracking

Originally, single-particle localization was limited to 2D dimensions, where only the x, y coordinates of an emitter are recovered [57]. Similar to works on depth from defocus, the depth of an emitter can be recovered from 2D measurements by considering a microscope's PSF. A standard microscope typically has a PSF resembling the circular Airy pattern; however, because it spreads out quickly its depth resolving range is limited. A few engineered PSFs—such as the double-helix PSF [46]—have since been proposed to improve the imaging range. In particular, Shechtman *et al*. finds the optimally informative PSF (dubbed the Fisher PSF) for a CMOS sensor to localize the 3D position of a single emitter [52]. A few other techniques for resolving the 3D location of particles have been proposed such as light-field-microscopy [37] and lensless imaging [35].

Unfortunately, these techniques are limited by the sub-kilohertz readout of conventional CMOS sensors. This hinders their use in imaging fast, dynamic processes such as blood flow [8] and voltage signals [1]. A few ultrafast imaging methods have also been proposed [15, 38, 63, 65] but require high-power illumination which can be phototoxic to certain organic samples. Recently, event cameras have been proposed as an alternative to CMOS sensors for 2D SMLM [9]. Another work proposes extending light-field-microscopy to event cameras to resolve 3D position but requires complex optical setups and sacrifices spatial resolution [20]. By designing optics to encode depth information into event streams, we can enable high-speed 3D SMLM.

2.3. Depth Estimation

Extracting 2D information from images tends to be a significantly easier task than extracting depth, hence, monocular depth estimation is often the bottleneck in 3D tracking performance. Structured light projectors [16] or time-of-flight sensors [13] use active illumination to extract depth information. Given these methods' reliance on an internal light source, performance can degrade in adverse lighting conditions. If we allow multiple views, stereo [21] or structure from motion [61] can triangulate 3D position. These methods are sensitive to occlusion and texture-less scenes and require multiple calibrated cameras. Many neural network approaches with all-in-focus CMOS images as input have been proposed [47, 59, 67, 68]. Recently, event-based depth estimation has made significant progress with neural networks [26, 42, 44, 53, 72]. Spiking neural networks have been proposed for spiking cameras, which similar to event cameras, offer asynchronous readout of pixels [69].

3. Theory

3.1. Event Camera Simulation

Let $(x(t), y(t), z(t))$ be the location of a point light source at time t. We focus on tracking points around some focal plane z, with $z(t) = z + \Delta z(t)$ and $z \gg |z(t)|$. In this context, a pin-hole camera would capture,

$$I_t(u, v) = \delta\left(u - f\frac{x(t)}{z + \Delta z(t)}, v - f\frac{y(t)}{z + \Delta z(t)}\right) \quad (1)$$

$$\approx \delta\left(u - \frac{f}{z}x(t), v - \frac{f}{z}y(t)\right) \quad (2)$$

where δ is the Dirac Delta function. Because f and z are constant, we will consider $x(t)$ and $y(t)$ pre-scaled for no-

tation sake. In practice, a camera captures a blurry image depending on the point-spread-function (PSF) it induces. A PSF h can be modeled with Fourier optics theory as a function of 3D-position x, y, z, amplitude modulation A caused by blocking light, and phase modulation ϕ^M caused by phase mask height variation [18].

$$h = \left| \mathcal{F} \left[A \exp \left(i\phi^{DF}(x,y,z) + i\phi^{M} \right) \right] \right|^2 \tag{3}$$

where $\phi^{DF}(x,y,z)$ is the defocus aberration due to the distance from the camera. Then, a point light source at location $(x(t), y(t), z(t))$ captured by a regular camera is

$$\begin{aligned} I^b_t(u,v) &= [h_{z(t)} * I_t](u,v) && (4) \\ &= h(x(t), y(t); z(t)). && (5) \end{aligned}$$

Note that because this PSF depends on depth, it can be used to encode depth information into I^b. Event cameras trigger events with respect to the log of photocurrent $L = \log(I^b)$ [14] where a pixel's photocurrent is linearly related to the wave intensity at that pixel. Specifically, an event is triggered when the absolute difference between the current intensity at $t + \tau$ and the reference intensity from t, $\Delta L(u,v) = L_{t+\tau}(u,v) - L_t(u,v)$, is greater than some threshold T.

$$O_t(u,v) = \begin{cases} +1 & \Delta L(u,v) > T \\ -1 & \Delta L(u,v) < -T \\ \text{none} & \text{otherwise} \end{cases} \tag{6}$$

In isolation, each event contains little information; however, a sequence of events can be highly informative [2, 31, 54]. Notably the inceptive event time-surfaces representation suggests the trailing events that occur after the first event correspond to the log-intensity change [6]. Therefore, by binning events over time, one can approximately recover the change in log intensity ΔL. Visuallly, we show the accumulated event frame approaches ΔL as the number of intermediate frames accumulated increases in Figure 2. We prove this approximation is at most off by 1 for an idealized event camera in Section S4 of the supplement. Therefore, our event measurement (6) can be simplified as,

$$O_t = \log\left(I^b_t\right) - \log\left(I^b_{t-\tau}\right). \tag{7}$$

3.2. Information

In the field of statistical information theory, the Fisher Information (FI) reports the amount of information gained about the parameters of a distribution, given a measurement. As such, we can use FI to express the effectiveness of PSFs at encoding depth information. The multi-parameter FI is represented as an $N \times N$ matrix where the i, j entry is defined as the variance of the score:

$$\mathcal{I}(\theta)_{i,j} = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta_i} \log f(X;\theta) \right) \left(\frac{\partial}{\partial \theta_j} \log f(X;\theta) \right) \mid \theta \right] \tag{8}$$

where θ is the set of parameters, θ_i is the ith parameter, and $f(X;\theta)$ is a probability density function for the distribution observation X is drawn from.

For traditional CMOS sensors, FI has been used to compare coded apertures and phase masks for a wide range of tasks such as depth estimation [45], hyper-spectral imaging [5], and detecting linear structures [17]. Those works have shown that the intrinsic photon shot noise in I^b can be modeled as a Poisson random variable with mean $\lambda = h(x,y,z)$. We derive the FI matrix for an event sensor.

Flashing light. As a warm-up, consider the SMLM technique for event cameras presented in [9], which assumes a blinking labeling model similar to STORM (stochastic optical reconstruction microscopy) [49], PALM (photoactivated localization microscopy) [7] and DNA-PAINT (DNA point accumulation for imaging in nano-scale topography) [51]. With this idealized model of an event camera, $\log I^b_{t-\tau} = 0$, so (7) reduces to

$$O_t = \log I^b_t. \tag{9}$$

By applying e^x to the measurement, we can indirectly measure I^b_t. Moreover, by applying standard results for FI of a Poisson distribution [28, 58], we can write the FI matrix for an event camera capturing a blinking particle as:

$$\mathcal{I}(\theta)_{i,j} = \sum_n^N \frac{1}{h(n) + \beta} \left(\frac{\partial h(n)}{\partial \theta_i} \right) \left(\frac{\partial h_z(n)}{\partial \theta_j} \right) \tag{10}$$

where N is the number of pixels, $h(n)$ is the PSF intensity at pixel n, β is background noise, and $\theta = \{x, y, z\}$ corresponds to the 3D location of a point source. Notice that this is the same result as in [45], suggesting that — in the context of blinking particles — the Fisher mask found in [52] for a traditional CMOS camera is also optimal for an event-based sensor.

Generalization. We now derive the positional information content for any event measurement. Rewriting (7) with logarithmic rules, we obtain,

$$O_t = \log \frac{I^b_t}{I^b_{t-\tau}}. \tag{11}$$

The inner expression is drawn from the ratio of Poisson random variables with means λ_t and $\lambda_{t-\tau}$. This can be approximated as a single Normal distribution [19]:

$$\frac{I^b_t}{I^b_{t-\tau}} \sim \mathcal{N}\left(\frac{\lambda_t}{\lambda_{t-\tau}}, \frac{\lambda_t}{\lambda^2_{t-\tau}} + \frac{\lambda^2_t}{\lambda^3_{t-\tau}} \right). \tag{12}$$

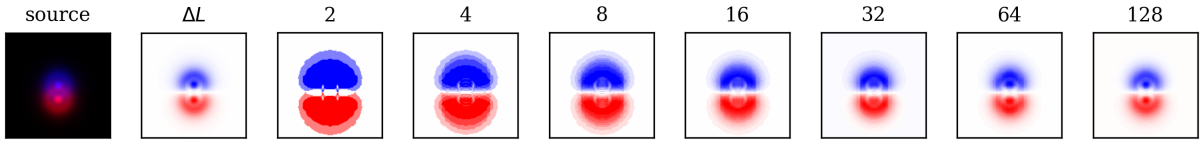


Figure 2. **Binning events approximates the log difference as the number of accumulated frames increases.** Consider a point source moving from the blue location to the red location at depth plane 1μm over a fixed time interval in the first image. The second image illustrates the direct access to the difference in (7), while the subsequent images demonstrate the effect of accumulating N event frames across the time interval. Observe how large N nearly recovers ΔL, demonstrating the validity of the approximation.

Similar to the flashing light example, we can exponentiate the measurement to recover this ratio. Using the symbolic mathematics solver SymPy [41], we evaluate the expectation in (10) with $\theta = \{x_t, y_t, z_t, x_{t-\tau}, y_{t-\tau}, z_{t-\tau}\}$ and $f(X;\theta)$ as the PDF of the normal distribution, yielding

$$\mathcal{I}(\theta) = \sum_{n}^{N} \frac{\mathcal{D}^T\mathcal{D}}{2\left(\mu+\nu\right)^2} \odot \begin{bmatrix} a & a & a & b & b & b \\ a & a & a & b & b & b \\ a & a & a & b & b & b \\ b & b & b & c & c & c \\ b & b & b & c & c & c \\ b & b & b & c & c & c \end{bmatrix} \tag{13}$$

where

$$\mu = \lambda_{t-\tau} = h(x_{t-\tau}, y_{t-\tau}, z_{t-\tau}) + \beta \tag{14}$$

$$\nu = \lambda_t = h(x_t, y_t, z_t) + \beta \tag{15}$$

$$\mu_i = \frac{\partial}{\partial\theta_i}\mu \tag{16}$$

$$\nu_i = \frac{\partial}{\partial\theta_i}\nu \tag{17}$$

$$\mathcal{D} = \begin{bmatrix} \mu_x/\mu & \mu_y/\mu & \mu_z/\mu & \nu_x/\nu & \nu_y/\nu & \nu_z/\nu \end{bmatrix} \tag{18}$$

$$a = 2\mu^2\nu + 4\mu^2 + 2\mu\nu^2 + 12\mu\nu + 9\nu^2 \tag{19}$$

$$b = -\left(2\mu^2\nu + 2\mu^2 + 2\mu\nu^2 + 7\mu\nu + 6\nu^2\right) \tag{20}$$

$$c = 2\mu^2\nu + \mu^2 + 2\mu\nu^2 + 4\mu\nu + 4\nu^2 \tag{21}$$

4. Method

4.1. Objective Function

Similar to existing work on 3D tracking for CMOS sensors, we can leverage the FI matrix to optimize optical parameters that efficiently encode depth information [52, 64]. Specifically, we compute the Cramér Rao Bound (CRB), which provides a fundamental bound on how accurately parameters can be estimated given a measurement. If $T(X)$ is the unbiased estimator for parameters θ, then the CRB is

$$CRB_i \equiv \left[\mathcal{I}(\theta)^{-1}\right]_i \leq \mathrm{cov}_\theta\left(T(X)\right)_i. \tag{22}$$

Then, the objective function we wish to minimize is

$$\mathcal{L}_{CRB} = \sum_{z\in Z}\sum_{i\in\theta}\sqrt{\left[\mathcal{I}(\theta)^{-1}\right]_{i,i}} \tag{23}$$

where Z is a set of depth planes.

4.2. Optical Parameter Representation

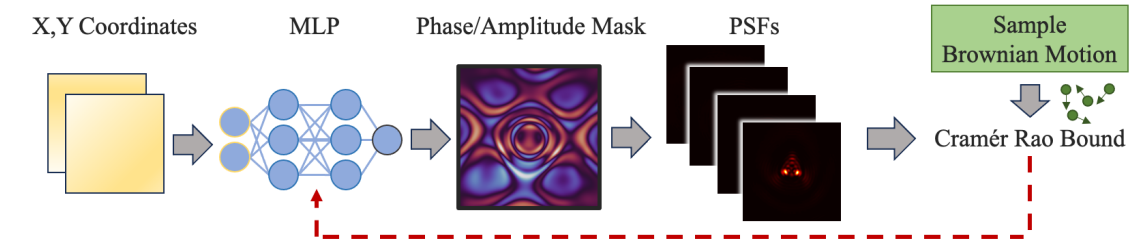
PSF manipulation is typically achieved through designed optical elements such as phase and amplitude masks. In general, phase masks are preferred over binary amplitude masks for their photon efficiency and continuous parametric representation, allowing for optimization via standard gradient descent methods. Inspired by [12], we demonstrate that implicit neural representations can model phase masks in such a way that results in more stable optimization and better-optimized mask designs. We use an architecture similar to the sinusoidal representation network (SIREN) presented in [56] to predict the phase delay caused by the mask at each location (u, v). Input data in $\mathbb{R}^2$ is processed by a four-layer multi-layer perceptron (MLP) with hidden feature size 128, and sin activation. We refer to this method as *Neural Phase Mask (NPM)*.

Phase masks offer many degrees of freedom and excellent light throughput, but can be relatively expensive to manufacture and are only effective for some frequencies. Meanwhile binary amplitude masks are cheap to manufacture (such as with consumer-grade 3D printers) and can operate across all frequencies (including x-ray), but offer fewer degrees of freedom.

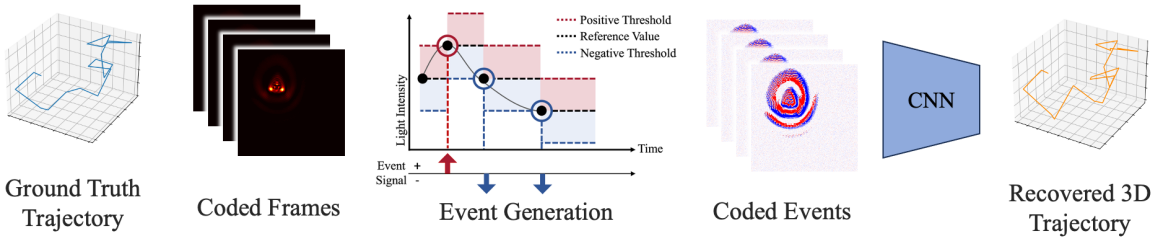
Historically, methods for designing optimal binary apertures have been fundamentally limited due to the lack of optimization techniques for discrete binary parameters. As a result, prior works [33, 70, 71] walk over a restricted search space, leaving ample room for improvement. To solve this issue, we propose a novel implicit neural representation for binary amplitude masks. We use an MLP to predict the percent of photons blocked at each mask location (u, v). The input in $\mathbb{R}^2$ is processed by a four-layer MLP with hidden feature size 128 and SoftPlus [43] activation. The output to the network is passed through a sigmoid. We refer to this method as *Neural Amplitude Mask (NAM)*.

5. Experimental Details

PSFs are simulated for a microscope imaging system with NA$=$ 1.4, index of refraction $n = 1.518$, wavelength $\lambda = 550$nm, magnification $M = 111.11$, 4f lens focal length $f = 150$mm, pixel pitch of 49.58μm, and resolution of 256×256. Each phase and amplitude mask is optimized



(a) Optical component optimization.



(b) Coded event 3D tracking

Figure 3. **System overview.** (a) An MLP produces a phase or amplitude mask based on a grid of x, y coordinates. The weights are updated through back-propagation of the CRB computed with Brownian Motion. (b) In simulation, coded events are generated by first rendering high-frame-rate coded CMOS frames and converting them to event frames. These measurements are passed to a 3D-tracking algorithm.

using $\mathcal{L}_{CRB}$ for 10,000 epochs. Because particle motion influences FI, we leverage Monte Carlo sampling while training to maximize information content for all motion directions. For each epoch, we compute the total CRB for 3 random orthogonal motions across 11 depth planes. We use the Adam [29] optimizer with parameters $\beta_1 = 0.99$, $\beta_2 = 0.999$, and a learning rate of 10^{-3}. Training and testing were conducted on NVIDIA RTX A5000 GPUs.

To validate our design's ability to track point sources, we train a Convolutional Neural Network (CNN) to map binned event frames to 3D locations. Events are accumulated over 16 refresh cycles to produce an accumulated event frame. These 256×256 single-channel images are processed by a CNN with 5 convolutional blocks and a linear output head. Each block is followed by batch normalization, ELU activation [11], and max pooling. The output is a normalized length 3 vector representing the position of the particle at a given time step. The CNN is trained on 3 Brownian motion trajectories. Each trajectory is sampled at 16,000 time steps. A 'coded' CMOS video frame is simulated by blurring a 300nm emitter with the optical component's PSF for the location and adding Gaussian noise (to simulate other noise sources such as thermal). Next, we generate a 'coded-event-stream' from the high-speed video using standard event camera simulator methods by tracking the per-pixel reference signal [23]. Finally, we bin every 16 frames to produce a 1000-frame 'coded-event-video'. The particle location at the end of the 16-frame bin is considered the ground truth position. We supplement this training with 2000 random starting positions and corresponding motion vectors. Each motion is scaled to have magnitude drawn from $\mathcal{N}(100\text{nm}, 20\text{nm})$. For each position-motion pair, we generate a 16 frame 'coded' CMOS video to accumulate into a 'coded-event-frame'. The CNN is trained for 100 epochs with the Adam optimizer. We also manufacture a lab prototype to the demonstrate practical benefits of coded apertures for event cameras (see Section S1 in the supplementary materials for details).

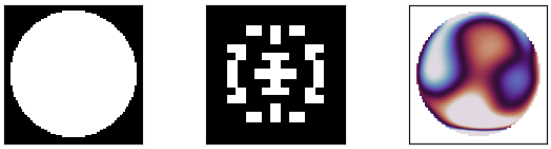


Figure 4. **Visualization of non-event camera-specific optical components.** Each component is placed in the same plane as a 150mm focal length lens.

6. Results

Because designed optics for event cameras is an emerging field, we compare our optimized phase and amplitude mask designs to components designed for traditional CMOS sensors: open aperture/Fresnel lens, Fisher phase mask [52] and Levin *et al.*'s amplitude mask [33] (Figure 4).

	Component	CRB (nm) ↓
	Open Aperture	80.8
Amplitude	Levin *et al.*	263.3
	NAM (Ours)	50.5
Phase	Fisher	36.3
	NPM (Ours)	**33.1**

Table 1. **Average CRB for each optical component** across a 3μm depth range for all 6 position parameters. Phase masks outperform amplitude masks due to higher light efficiency, and our neural-designed phase mask is best.

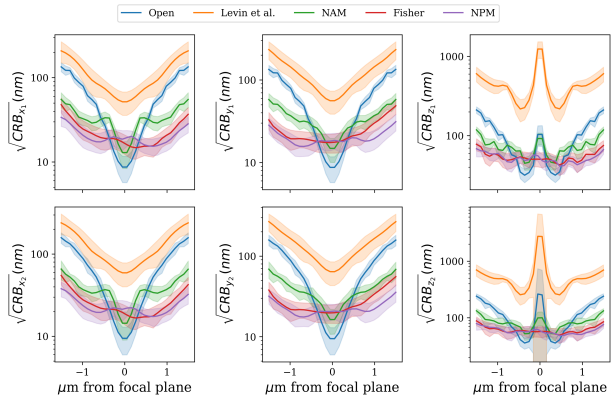


Figure 5. **3D localization CRB with respect to depth**. First row: particle's x, y, z position at time $t - \tau$. Second row: particle's x, y, z position at time t. Observe the bound increases as the source drifts from the focal plane.

6.1. Cramér Rao Bound

We simulate Brownian motion by sampling 1000 unit direction vectors and independently scaling them by a magnitude drawn from $\mathcal{N}(100\text{nm}, 20\text{nm})$. The speed is relative to the event camera refresh rate, with a 1000 accumulated-event-frame per second system, this motion simulates a range of biological processes such as molecular diffusion [66]. We then evaluate the average CRB over the 1000 motions at 30 depth planes spaced evenly on a 3μm range around the focal plane. For all 6 position parameters, we plot the CRB trend with respect to depth (Figure 5). Observe that each optical system performs worse as a point source moves away from the focal plane as the defocus change decreases. Although an open-aperture lens is slightly better around the focal plane, its bound increases at a higher rate than the other designs. We also report the average CRB over all parameters and depth slices to demonstrate our neural-based phase mask is best overall (Table 1).

	Component	RMSE (nm) ↓ 3D	L_1 (nm) ↓ z
	Open Aperture	617	936
Amplitude	Levin *et al.*	764	1036
	NAM (Ours)	66.0	49.2
Phase	Fisher	52.6	44.2
	NPM (Ours)	**51.2**	**39.2**

Table 2. **Tracking accuracy comparison.** We present quantitative results on 3D trajectory recovery for known optical designs. Our event CRB loss function found the best-performing design. Although only slightly improved in overall 3D tracking, our design noticeably improves depth recovery.

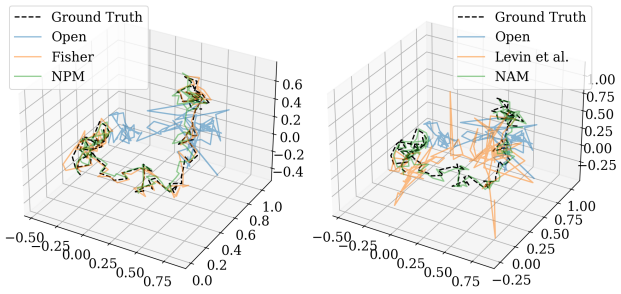


Figure 6. **Recovered 3D position over Brownian motion sequence with coded event frames.** Left: phase mask methods, right: amplitude mask methods. Observe trajectories reconstructed from phase mask-coded events more closely align with ground-truth positions. Units in microns.

6.2. 3D Tracking

We validate our theoretical results in simulation by tracking a 3D moving emitter across a $8\mu\text{m} \times 8\mu\text{m} \times 4\mu\text{m}$ volume. After training a CNN to decode 3D position from coded event frames, we evaluate our network tracking performance on 5 sequences of Brownian motion, each consisting of 1000 binned frames. Table 2 shows our event camera-specific optical designs minimize 3D tracking error more than conventional designs. Additionally, our method is substantially better at depth plane recovery. Qualitative results in Figure 6 demonstrate that 3D positions recovered using our designs more tightly fit ground-truth trajectories.

7. Ablation Studies

7.1. Optical Representations

Additionally, we compare 3D tracking results using two different amplitude mask representations: pixel-wise and neural amplitude mask (Figure 7) and three different phase mask representations: pixel-wise, Zernike basis, and neural phase mask (Figure 8). As shown in Table 3, our implicit

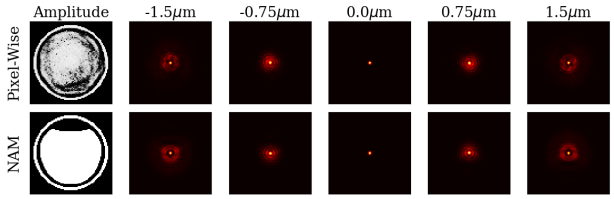


Figure 7. **Designed amplitude masks and corresponding PSFs.** Top: pixel-wise representation. Bottom: implicit neural representation.

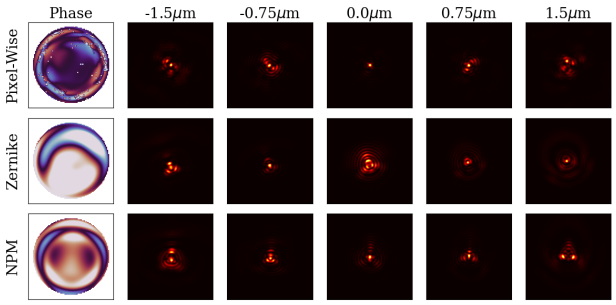


Figure 8. **Designed phase masks and corresponding PSFs.** Top: pixel-wise representation. Middle: first 55 Zernike coefficients representation. Bottom: implicit neural representation.

	Representation	CRB (nm) ↓
Amplitude	Pixel-Wise	65.5
	NAM	50.5
Phase	Pixel-Wise	34.2
	Zernike	34.8
	NPM	**33.1**

Table 3. **Average CRB of different optimized representations across a 3μm depth range**. Notice the neural representations outperform their pixel-wise counterparts.

neural representation-based methods achieve a lower average error bound than alternative representations, despite being two times smaller than pixel-wise representations with respect to the number of parameters. As expected, phase mask results generally outperform the amplitude mask results (Figure 9). However, our novel neural binary aperture makes optimizing amplitude masks more tractable. We observe that pixel-wise representations not only yield difficult-to-manufacture apertures but also suboptimal performance. In terms of 3D tracking, the implicit neural representations produce a smaller error on average (Table 4) and more accurately match sampled 3D trajectories (Figure 10).

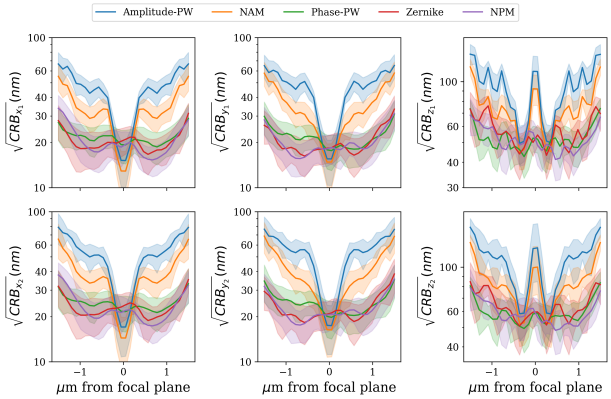


Figure 9. **Effect of optical parameterization on 3D localization CRB**. First row: particle's x, y, z position at time $t - \tau$. Second row: particle's x, y, z position at time t. Our implicit neural representations are particularly advantageous for amplitude masks.

		RMSE (nm) ↓	L_1 (nm) ↓
	Component	3D	z
Amplitude	Pixel-Wise	120	103
	NAM	66.0	49.2
Phase	Pixel-Wise	56.5	45.9
	Zernike	51.3	50.2
	Our NPM	**51.2**	**39.2**

Table 4. **Effect of optimized mask parameterization on tracking accuracy.** Average distance between ground-truth Brownian motion and the recovered 3D position is minimized with our neural-based designs.

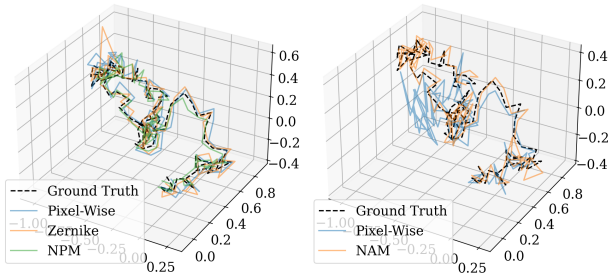


Figure 10. **Effect of optical representation on 3D trajectory recovery.** Left: phase mask methods, right: amplitude mask methods. Observe that neural representations produce tighter reconstructions. Units in microns.

7.2. Tracking Limits

In this section, we explore the limits of 3D tracking with variable external factors. For each experiment, we compute the average CRB over 30 depth slices and 6 parameters for 3 orthogonal unit directions (x, y, and z). First, as the number

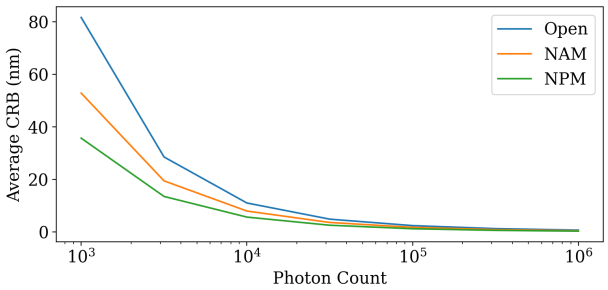


Figure 11. **Flux effect on CRB.** With more available photons, the signal-to-noise ratio increases, so the 3D information content is more reliable, and the bound on 3D tracking error decreases.

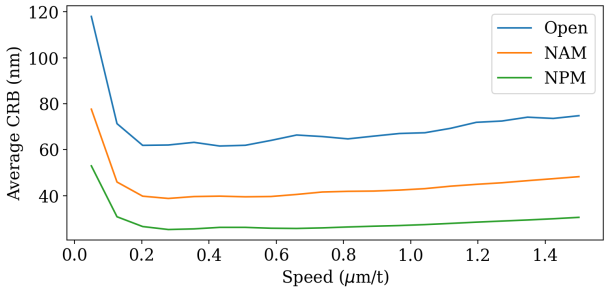


Figure 12. **Speed effect on CRB.** Too-slow moving particles trigger fewer events yielding a worse CRB. Similarly, as a particle moves faster the delay between triggers leads to fewer events.

of available photons increases, the lower bound on 3D position estimation monotonically decreases (Figure 11). More available photons equate to a higher signal-to-noise ratio. Additionally, this result helps explain why phase masks outperform amplitude masks. Second, we show extremely slow-moving particles (less than nanometers per refresh rate) experience a significantly higher CRB (Figure 12). Minimal movement indicates smaller intensity changes and thus an event camera would trigger fewer events. On the other side, as a particle moves faster, the number of events will decrease as there is a non-zero delay between when an event camera can trigger sequential events. Our learned phase mask is more robust to speed changes than an open aperture and our learned amplitude mask. Third, when the percentage of photons due to background noise increases, the bound on error also increases (Figure 13). We design our masks with 1% of captured photons attributable to the background, but the learned designs are more resistant to degraded conditions than an open aperture.

We also explore the effect of modifying the accumulation period in Section S2 and how the optimal design changes with respect to speed in Section S3 of the supplement.

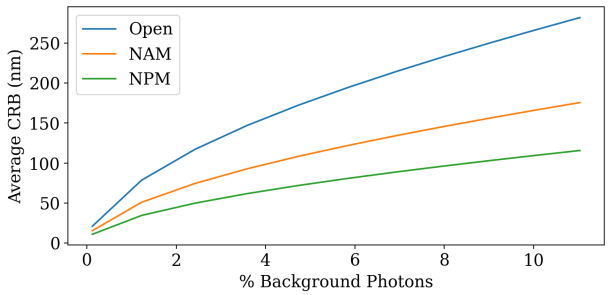


Figure 13. **Background photon effect on CRB.** As the percentage of photons hitting the sensor due to background noise increases, CRB also increases. The impact is minimal in our method.

8. Limitations

While we were successful in designing optics to improve performance on 3D tracking with event cameras, our method carries some limitations. First, although our binned event frames can be obtained at kHz refresh rates, they do not take full advantage of the asynchronous nature of event cameras. Second, our bounds are for an idealized event camera model with no read-noise. It would be impossible to outperform these bounds, but there might exist a tighter bound that accounts for these hardware imperfections. Lastly, we only consider single-emitter images. With multiple point sources, the resolving accuracy between single points may be more limited.

9. Conclusion

This work introduces PSF-engineering to neuromorphic event-based sensors. We first derive information theoretical limits on 3D point localization and tracking. We demonstrate that existing amplitude and phase mask designs are suboptimal for tracking moving emitters and design new optical elements for this task. Additionally, to overcome the non-convexity of this optimization problem, we introduce a novel implicit neural representation for optical components. Finally, we validate the effectiveness of our designs in simulation and compare against state-of-the-art mask designs. Our work unlocks not only highly performant optics for event cameras but also the ability to design highly expressive elements for other sensors.

Acknowledgements

This work was supported in part by the Joint Directed Energy Transition Office, AFOSR Young Investigator Program award no. FA9550-22-1-0208, ONR award no. N00014-23-1-2752 and N00014-17-1-2622, Dolby Labs, SAAB, Inc, and National Science Foundation grants BCS 1824198 and CNS 1544787. The support of the Maryland Robotics Center under a postdoctoral fellowship to C.S., is also gratefully acknowledged.

References

[1] Ahmed S. Abdelfattah, Jihong Zheng, Amrita Singh, Yi-Chieh Huang, Daniel Reep, Getahun Tsegaye, Arthur Tsang, Benjamin J. Arthur, Monika Rehorova, Carl V. L. Olson, Yichun Shuai, Lixia Zhang, Tian-Ming Fu, Daniel E. Milkie, Maria V. Moya, Timothy D. Weber, Andrew L. Lemire, Christopher A. Baker, Natalie Falco, Qinsi Zheng, Jonathan B. Grimm, Mighten C. Yip, Deepika Walpita, Martin Chase, Luke Campagnola, Gabe J. Murphy, Allan M. Wong, Craig R. Forest, Jerome Mertz, Michael N. Economo, Glenn C. Turner, Minoru Koyama, Bei-Jung Lin, Eric Betzig, Ondrej Novak, Luke D. Lavis, Karel Svoboda, Wyatt Korff, Tsai-Wen Chen, Eric R. Schreiter, Jeremy P. Hasseman, and Ilya Kolb. Sensitivity optimization of a rhodopsin-based fluorescent voltage indicator. *Neuron*, 111(10):1547–1563.e9, 2023. 2

[2] Ignacio Alzugaray and Margarita Chli. Asynchronous corner detection and tracking for event cameras in real time. *IEEE Robotics and Automation Letters*, 3(4):3177–3184, 2018. 3

[3] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, Jeff Kusnitz, Michael Debole, Steve Esser, Tobi Delbruck, Myron Flickner, and Dharmendra Modha. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[4] Anastasios N. Angelopoulos, Julien N.P. Martel, Amit P.S. Kohli, Jorg Conradt, and Gordon Wetzstein. Event based, near-eye gaze tracking beyond 10,000 hz. *IEEE Transactions on Visualization and Computer Graphics (Proc. VR)*, 2021. 1

[5] Seung-Hwan Baek, Hayato Ikoma, Daniel S. Jeon, Yuqi Li, Wolfgang Heidrich, Gordon Wetzstein, and Min H. Kim. Single-shot hyperspectral-depth imaging with learned diffractive optics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2651–2660, 2021. 3

[6] R. Wes Baldwin, Mohammed Almatrafi, Jason R. Kaufman, Vijayan Asari, and Keigo Hirakawa. Inceptive event time-surfaces for object classification using neuromorphic cameras. In *Image Analysis and Recognition - 16th International Conference, ICIAR 2019, Proceedings*, pages 395–403, Germany, 2019. Springer Verlag. 3

[7] Eric Betzig, George H. Patterson, Rachid Sougrat, O. Wolf Lindwasser, Scott Olenych, Juan S. Bonifacino, Michael W. Davidson, Jennifer Lippincott-Schwartz, and Harald F. Hess. Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, 313(5793):1642–1645, 2006. 3

[8] Matthew B Bouchard, Brenda R Chen, Sean A Burgess, and Elizabeth M C Hillman. Ultra-fast multispectral optical imaging of cortical oxygenation, blood flow, and intracellular calcium dynamics. *Opt Express*, 17(18):15670–15678, 2009. 2

[9] Clément Cabriel, Tual Monfort, Christian G. Specht, and Ignacio Izeddin. Event-based vision sensor for fast and dense single-molecule localization microscopy. *Nature Photonics*, 2023. 1, 2, 3

[10] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *Proc. IEEE ICCV*, 2019. 2

[11] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus), 2016. 5

[12] Brandon Y. Feng, Haiyun Guo, Mingyang Xie, Vivek Boominathan, Manoj K. Sharma, Ashok Veeraraghavan, and Christopher A. Metzler. Neuws: Neural wavefront shaping for guidestar-free imaging through static and dynamic scattering media. *Science Advances*, 9(26):eadg4671, 2023. 4

[13] Sergi Foix, Guillem Alenya, and Carme Torras. Lock-in time-of-flight (tof) cameras: A survey. *IEEE Sensors Journal*, 11(9):1917–1926, 2011. 2

[14] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 1, 3

[15] Liang Gao, Jinyang Liang, Chiye Li, and Lihong V. Wang. Single-shot compressed ultrafast photography at one hundred billion frames per second. *Nature*, 516(7529):74–77, 2014. 2

[16] Jason Geng. Structured-light 3d surface imaging: a tutorial. *Adv. Opt. Photon.*, 3(2):128–160, 2011. 2

[17] Bhargav Ghanekar, Vishwanath Saragadam, Dushyant Mehra, Anna-Karin Gustavsson, Aswin C. Sankaranarayanan, and Ashok Veeraraghavan. Ps ^{2} f: Polarized spiral point spread function for single-shot 3d sensing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–12, 2022. 3

[18] Joseph W. Goodman. *Introduction to fourier optics*. Freeman, 2017. 3

[19] Tralissa F Griffin. Distribution of the ratio of two poisson random variables. Master's thesis, Texas Tech University, 1992. 3

[20] Ruipeng Guo, Qianwan Yang, Andrew S. Chang, Guorong Hu, Joseph Greene, Christopher V. Gabel, Sixian You, and Lei Tian. Eventlfm: Event camera integrated fourier light field microscopy for ultrafast 3d imaging, 2023. 2

[21] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 2

[22] Carlos Hinojosa, Juan Carlos Niebles, and Henry Arguello. Learning privacy-preserving optics for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2573–2582, 2021. 2

[23] Y Hu, S C Liu, and T Delbruck. v2e: From video frames to realistic DVS events. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2021. 5

[24] Craig Iaboni, Himanshu Patel, Deepan Lobo, Ji-Won Choi, and Pramod Abichandani. Event camera based real-time de-

tection and tracking of indoor ground robots. *IEEE Access*, 9:166588–166602, 2021. 1

[25] Hayato Ikoma, Cindy M. Nguyen, Christopher A. Metzler, Yifan Peng, and Gordon Wetzstein. Depth from defocus with learned optics for imaging and occlusion-aware depth estimation. *IEEE International Conference on Computational Photography (ICCP)*, 2021. 2

[26] Daniel Gehrig Javier Hidalgo-Carrio and Davide Scaramuzza. Learning monocular dense depth from events. *IEEE International Conference on 3D Vision.(3DV)*, 2020. 2

[27] James M. Jusuf and Matthew D. Lew. Towards optimal point spread function design for resolving closely spaced emitters in three dimensions. *Opt. Express*, 30(20):37154–37174, 2022. 2

[28] Steven M. Kay. *Fundamentals of Statistical Signal Processing*. Prentice-Hall, 1 edition, 1993. 3

[29] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015. 5

[30] Anika Kinkhabwala, Zongfu Yu, Shanhui Fan, Yuri Avlasevich, Klaus Müllen, and W. E. Moerner. Large single-molecule fluorescence enhancements produced by a bowtie nanoantenna. *Nature Photonics*, 3(11):654–657, 2009. 1

[31] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE Trans Pattern Anal Mach Intell*, 39(7):1346–1359, 2017. 3

[32] Jun Haeng Lee, Tobi Delbruck, Michael Pfeiffer, Paul K J Park, Chang-Woo Shin, Hyunsurk Eric Ryu, and Byung Chang Kang. Real-time gesture interface based on event-driven processing from stereo silicon retinas. *IEEE Trans Neural Netw Learn Syst*, 25(12):2250–2263, 2014. 1

[33] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)*, 26 (3):70–es, 2007. 2, 4, 5

[34] Lingen Li, Lizhi Wang, Weitao Song, Lei Zhang, Zhiwei Xiong, and Hua Huang. Quantization-aware deep optics for diffractive snapshot hyperspectral imaging. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19748–19757, 2022. 2

[35] Fanglin Linda Liu, Grace Kuo, Nick Antipa, Kyrollos Yanny, and Laura Waller. Fourier diffuserscope: single-shot 3d fourier light field microscopy with a diffuser. *Opt. Express*, 28(20):28969–28986, 2020. 2

[36] Xin Liu, Linpei Li, Xu Liu, Xiang Hao, and Yifan Peng. Investigating deep optics model representation in affecting resolved all-in-focus image quality and depth estimation fidelity. *Opt. Express*, 30(20):36973–36984, 2022. 2

[37] A. Llavador, J. Sola-Pikabea, G. Saavedra, B. Javidi, and M. Martínez-Corral. Resolution improvements in integral microscopy with fourier plane recording. *Opt. Express*, 24(18): 20792–20798, 2016. 2

[38] Yayao Ma, Youngjae Lee, Catherine Best-Popescu, and Liang Gao. High-speed compressed-sensing fluorescence lifetime imaging microscopy of live cells. *Proceedings of the National Academy of Sciences*, 118(3):e2004176118, 2021. 2

[39] Stephanie A Maynard, Philippe Rostaing, Natascha Schaefer, Olivier Gemin, Adrien Candat, Andréa Dumoulin, Carmen Villmann, Antoine Triller, and Christian G Specht. Identification of a stereotypic molecular arrangement of endogenous glycine receptors at spinal cord synapses. *eLife*, 10:e74441, 2021. 1

[40] Christopher A. Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[41] Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, 2017. 4

[42] Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Event-intensity stereo: Estimating depth by the best of both worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4258–4267, 2021. 2

[43] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, page 807–814, Madison, WI, USA, 2010. Omnipress. 4

[44] Yeongwoo Nam, Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Stereo depth from events cameras: Concentrate and focus on the future. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6114–6123, 2022. 2

[45] Raimund J Ober, Sripad Ram, and E Sally Ward. Localization accuracy in single-molecule microscopy. *Biophysical journal*, 86(2):1185–1200, 2004. 3

[46] Sri Rama Prasanna Pavani, Michael A. Thompson, Julie S. Biteen, Samuel J. Lord, Na Liu, Robert J. Twieg, Rafael Piestun, and W. E. Moerner. Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. *Proceedings of the National Academy of Sciences*, 106(9):2995–2999, 2009. 2

[47] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 2

[48] Juan Pablo Rodríguez-Gómez, Raul Tapia, Maria del Mar Guzmán Garcia, Jose Ramiro Martínez-de Dios, and Anibal Ollero. Free as a bird: Event-based dynamic sense-and-avoid for ornithopter robot flight. *IEEE Robotics and Automation Letters*, 7(2):5413–5420, 2022. 1

[49] Michael J Rust, Mark Bates, and Xiaowei Zhuang. Sub-diffraction-limit imaging by stochastic optical reconstruction

microscopy (storm). *Nature Methods*, 3(10):793–796, 2006. 3

[50] Sachin Shah, Sakshum Kulshrestha, and Christopher A. Metzler. Tidy-psfs: Computational imaging with time-averaged dynamic point-spread-functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10657–10667, 2023. 2

[51] Alexey Sharonov and Robin M. Hochstrasser. Wide-field subdiffraction imaging by accumulated binding of diffusing probes. *Proceedings of the National Academy of Sciences*, 103(50):18911–18916, 2006. 3

[52] Yoav Shechtman, Steffen J. Sahl, Adam S. Backer, and W. E. Moerner. Optimal point spread function design for 3d imaging. *Phys. Rev. Lett.*, 113:133902, 2014. 1, 2, 3, 4, 5

[53] Peilun Shi, Jiachuan Peng, Jianing Qiu, Xinwei Ju, Frank Po Wen Lo, and Benny Lo. Even: An event-based framework for monocular depth estimation at adverse night conditions, 2023. 2

[54] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[55] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics (TOG)*, 37(4):114, 2018. 2

[56] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Proc. NeurIPS*, 2020. 4

[57] Alex Small and Shane Stahlheber. Fluorophore localization algorithms for super-resolution microscopy. *Nature Methods*, 11(3):267–279, 2014. 2

[58] Donald L. Snyder and Michael I. Miller. *Random Point Processes in time and space*. Springer, 2 edition, 1991. 3

[59] Jaime Spencer, Richard Bowden, and Simon Hadfield. Defeat-net: General monocular depth via simultaneous unsupervised representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[60] Luc Tinch, Nitesh Menon, Keigo Hirakawa, and Scott McCloskey. Event-based detection, tracking, and recognition of unresolved moving objects. *Advanced Maui Optical and Space Surveillance Technologies (AMOS) Conference*, 2022. 1

[61] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992. 2

[62] Hippolyte Verdier, François Laurent, Alhassan Cassé, Christian L. Vestergaard, Christian G. Specht, and Jean-Baptiste Masson. A maximum mean discrepancy approach reveals subtle changes in α-synuclein dynamics. *bioRxiv*, 2022. 1

[63] Jianglai Wu, Yajie Liang, Shuo Chen, Ching-Lung Hsu, Mariya Chavarha, Stephen W Evans, Dongqing Shi, Michael Z Lin, Kevin K Tsia, and Na Ji. Kilohertz two-photon fluorescence microscopy imaging of neural activity in vivo. *Nat Methods*, 17(3):287–290, 2020. 2

[64] Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. Phasecam3d — learning phase masks for passive single view depth estimation. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12, 2019. 2, 4

[65] Sheng Xiao, John T. Giblin, David A. Boas, and Jerome Mertz. High-throughput deep tissue two-photon microscopy at kilohertz frame rates. *Optica*, 10(6):763–769, 2023. 2

[66] Nicole J Yang and Marlon J Hinner. Getting across the cell membrane: an overview for small molecules, peptides, and proteins. *Methods Mol Biol*, 1266:29–53, 2015. 6

[67] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9043–9053, 2023. 2

[68] Zunzhi You, Yi-Hsuan Tsai, Wei-Chen Chiu, and Guanbin Li. Towards interpretable deep networks for monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12879–12888, 2021. 2

[69] Jiyuan Zhang, Lulu Tang, Zhaofei Yu, Jiwen Lu, and Tiejun Huang. Spike transformer: Monocular depth estimation for spiking camera. In *Computer Vision – ECCV 2022*, pages 34–52, Cham, 2022. Springer Nature Switzerland. 2

[70] Changyin Zhou and Shree Nayar. What are good apertures for defocus deblurring? In *2009 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8, 2009. 4

[71] Changyin Zhou, Stephen Lin, and Shree Nayar. Coded aperture pairs for depth from defocus. In *2009 IEEE 12th International Conference on Computer Vision*, pages 325–332, 2009. 4

[72] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2