

Computational Imaging with Event Cameras

Chris Metzler



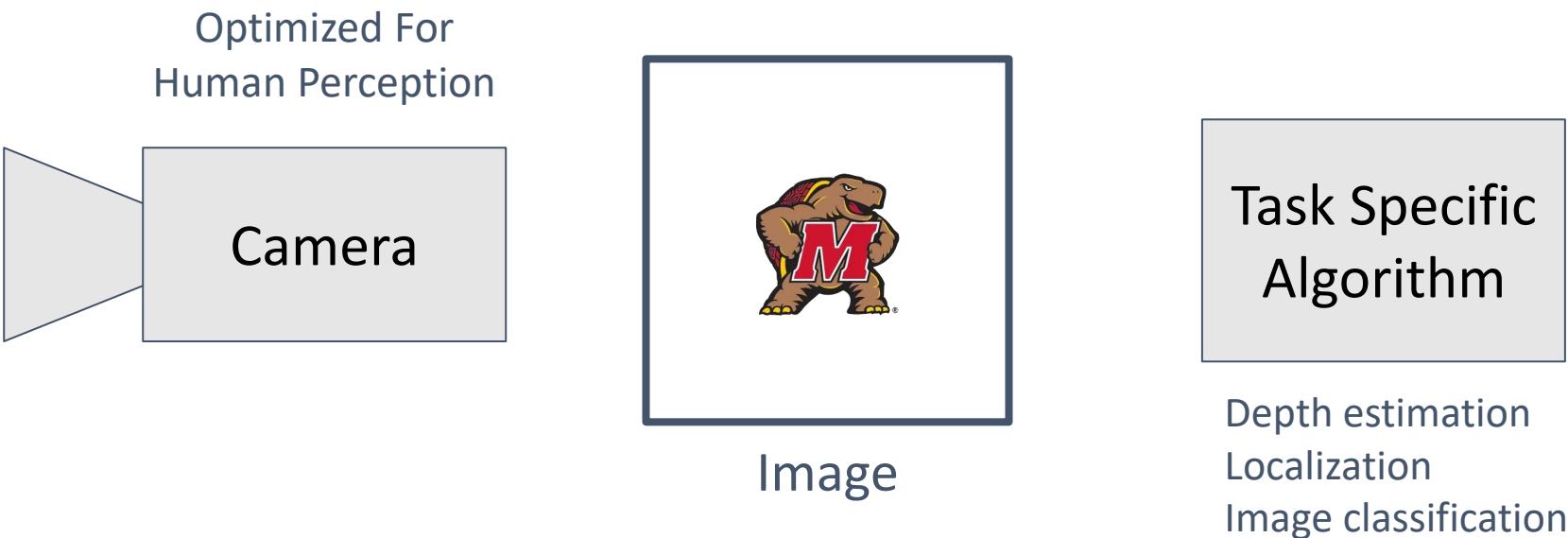
UNIVERSITY OF
MARYLAND



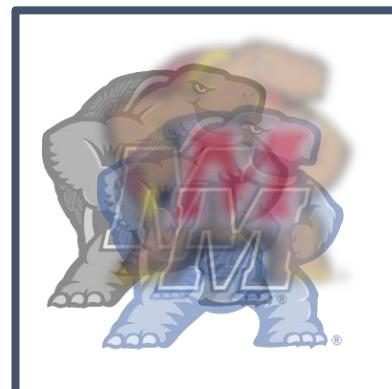
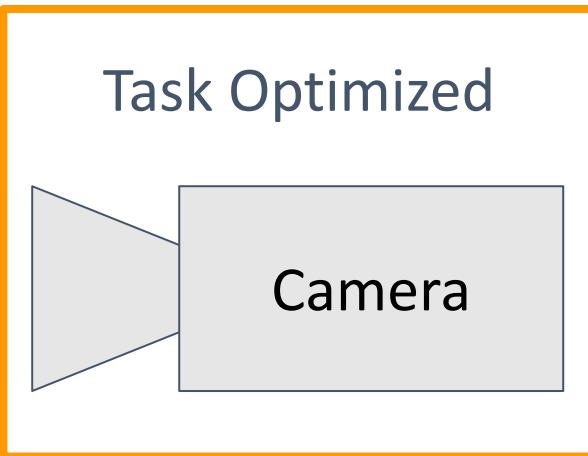
UMD Intelligent
Sensing Lab

*Computational Imaging is the
Co-Design of Optics and Algorithms*

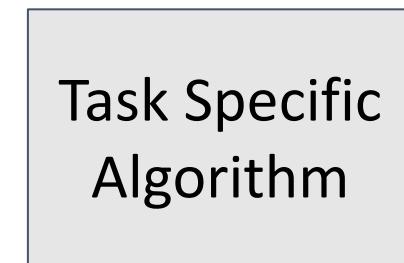
Traditional Computer Vision



Computational Imaging



Coded Image



Depth estimation
Localization
Image classification

Computational Imaging in Nature



Polarization



Multi-focus



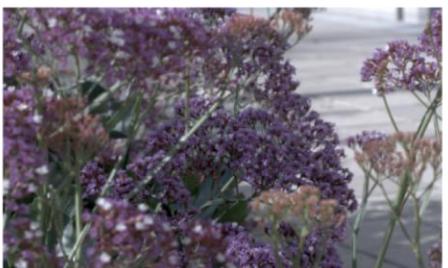
High dynamic range



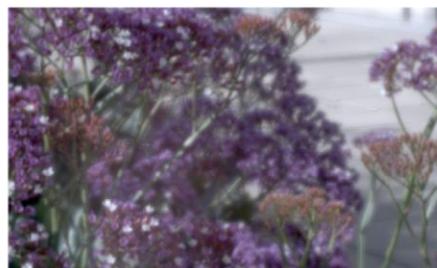
Ultra-violet

Depth Estimation with Computational Imaging

Conventional



Coded Image

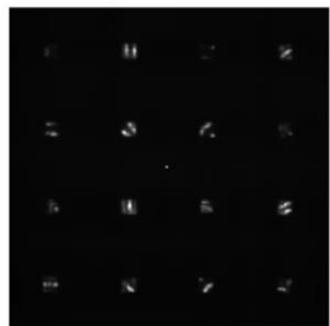


Depth Estimation with Computational Imaging



Computational Imaging for ...

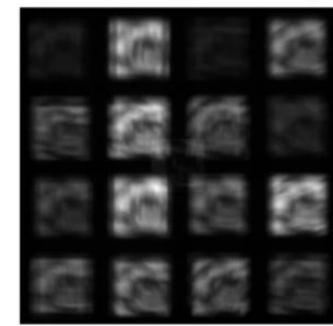
captured PSF



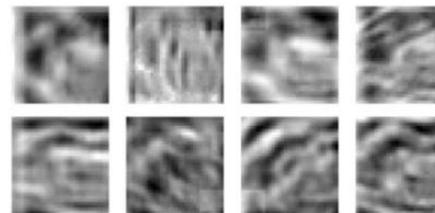
captured input image



captured sensor image



pseudonegative sub-images,
from capture



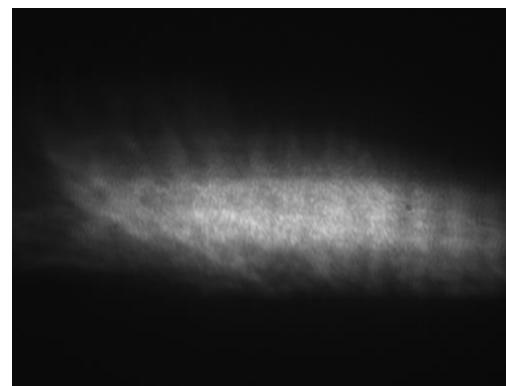
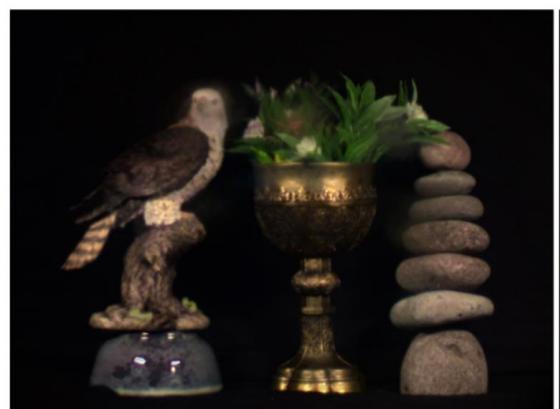
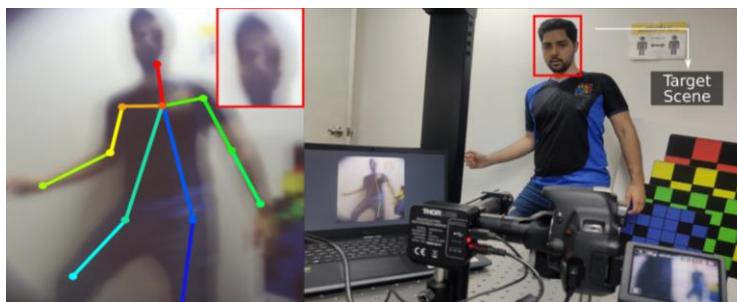
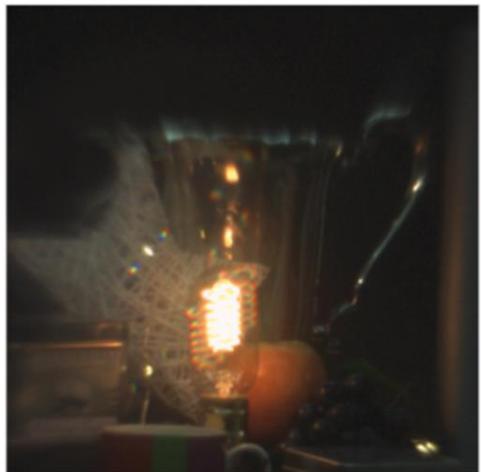
test accuracy: 44.40%

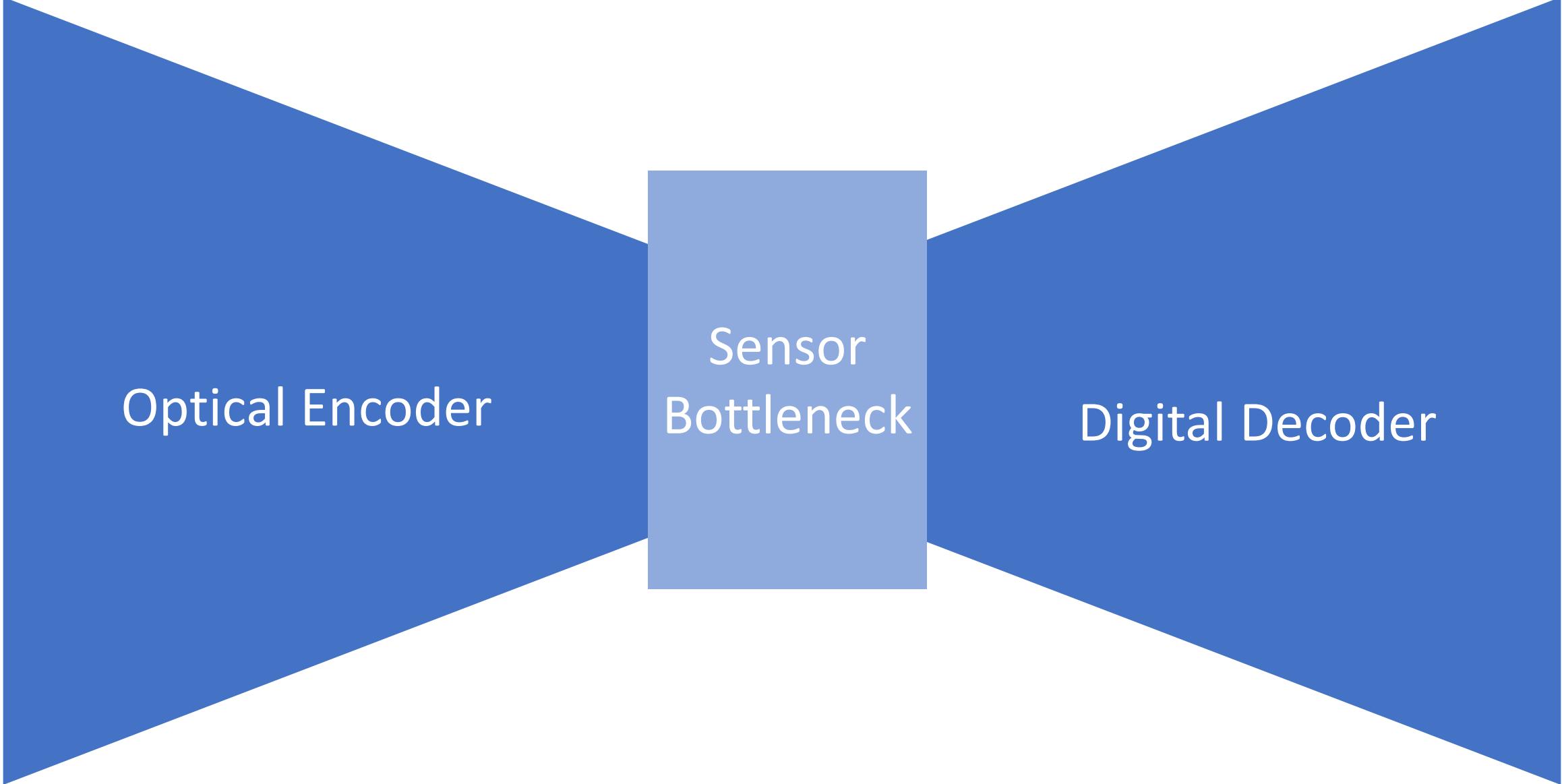
Image classification
(Chang et al. 2018)

High dynamic range
(M. et al. 2020)

Privacy preservation
(Hinojosa et al. 2021)

Seeing through obstructions
(Shi et al. 2022, Xie et al. 2024)





Optical Encoder

Sensor
Bottleneck

Digital Decoder

Todays Talk:

Part 1
Optical Encoder

Sensor
Bottleneck

Part 2
Digital Decoder

Todays Talk:

Part 1
Optical Encoder

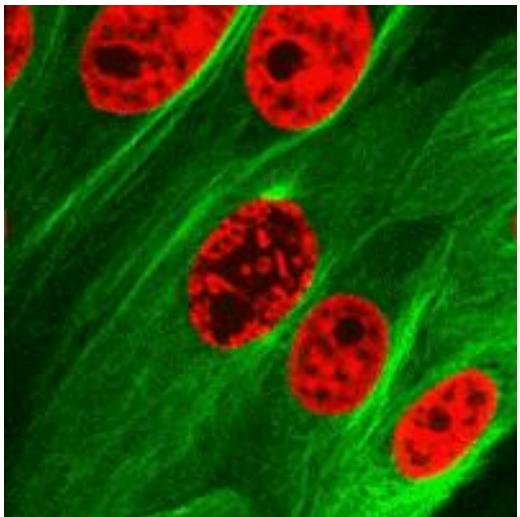
Sensor
Bottleneck

Part 2
Digital Decoder

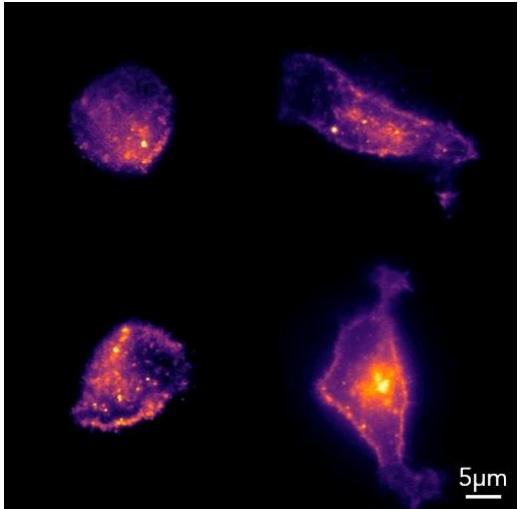
Sachin Shah



Our Goal: Passive 3D Sensing at 1000+ fps



[Agarwal et al. 2018]



[Knowles and Mahmood]



[National Geography]



[DJI]



[US Navy 2014]



[PBS 2016]



[SpaceX 2019]

One Option: Stereo Event Camera Systems



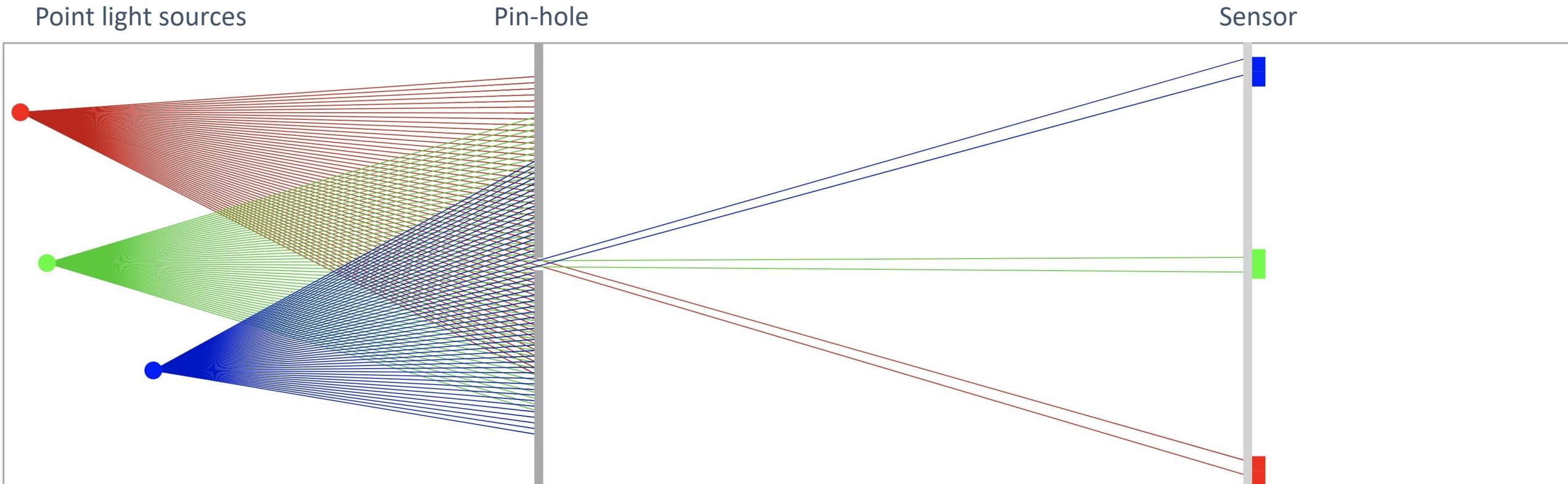
- Expensive
- Hard to synchronize
- Bulky

Alternative: Event Camera with Coded Optics

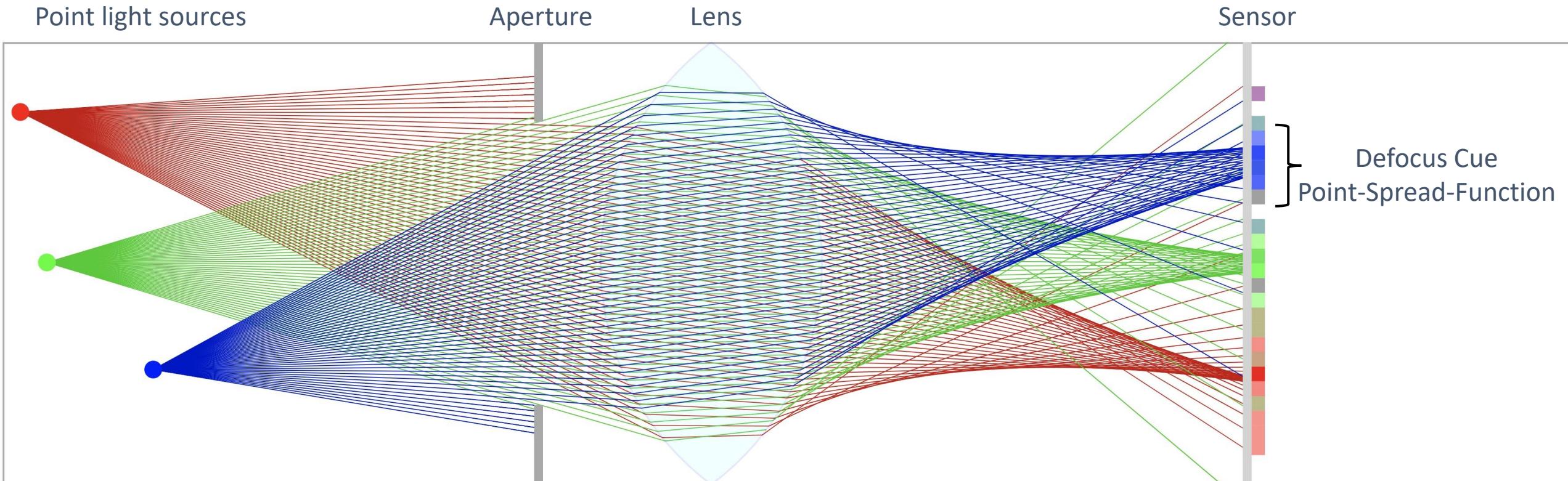


- ~~Expensive~~
- ~~Hard to synchronize~~
- ~~Bulky~~

How a pinhole camera works

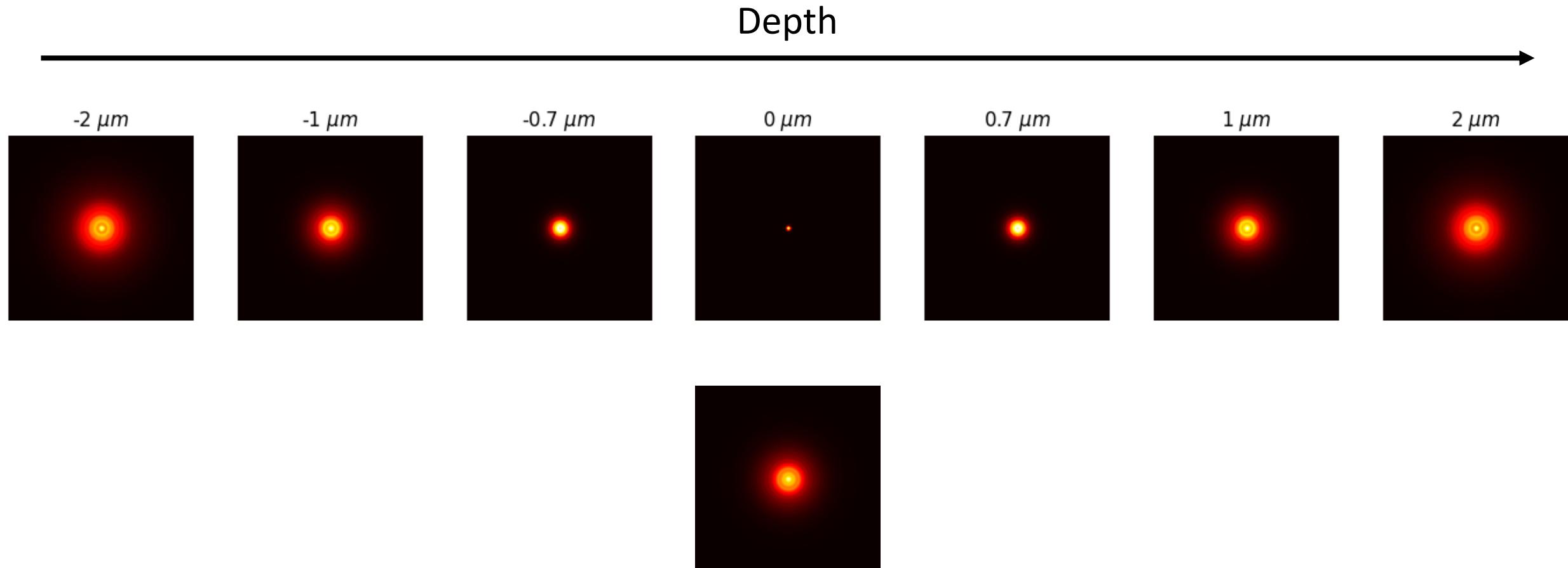


How a real camera works



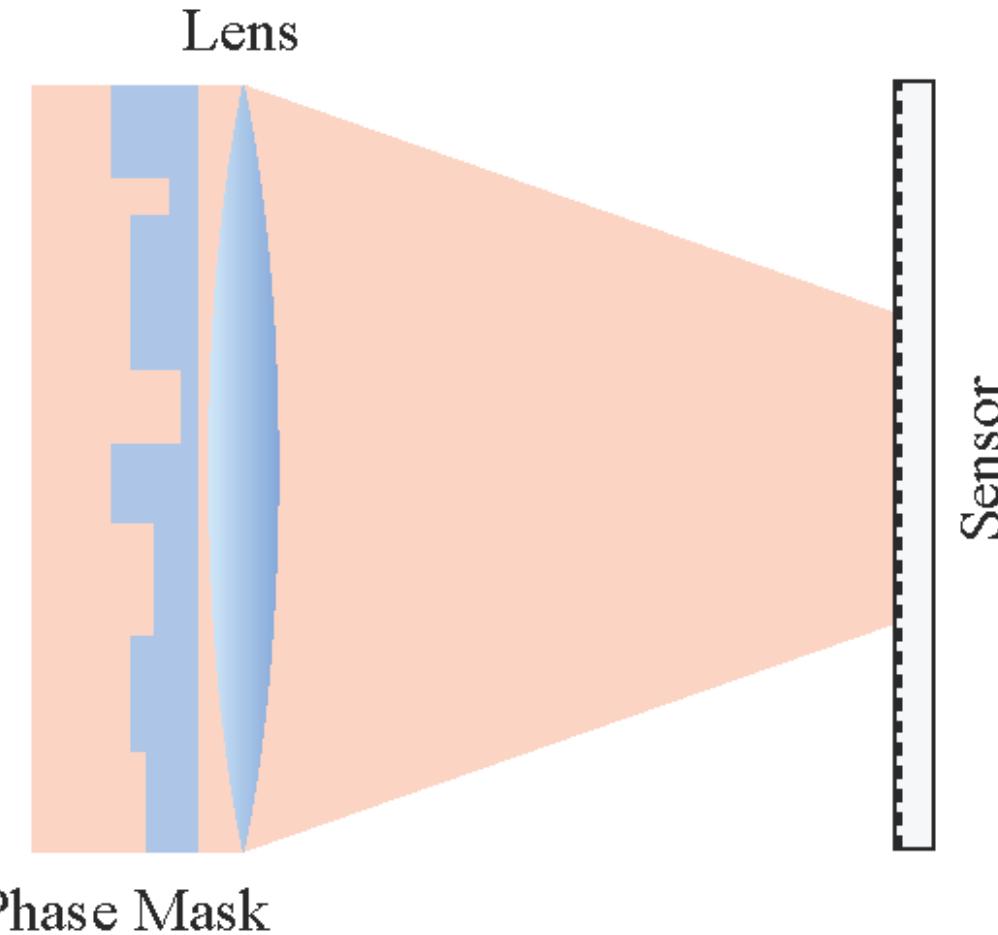
Cameras naturally encode some depth information

Estimating Depth From Defocus Cues



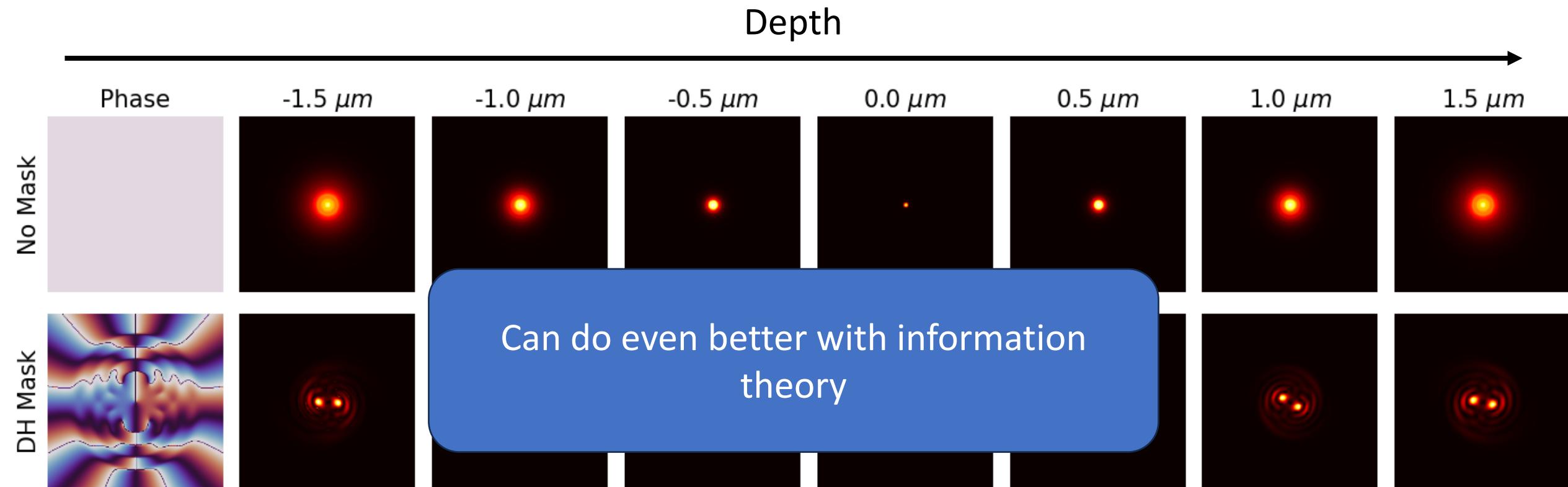
Depth is challenging to estimate with conventional optical systems

By introducing a phase mask into the optics, we can change the shape of the PSF



We can use elaborate defocus cues to “encode” depth into the images

Double-Helix PSF

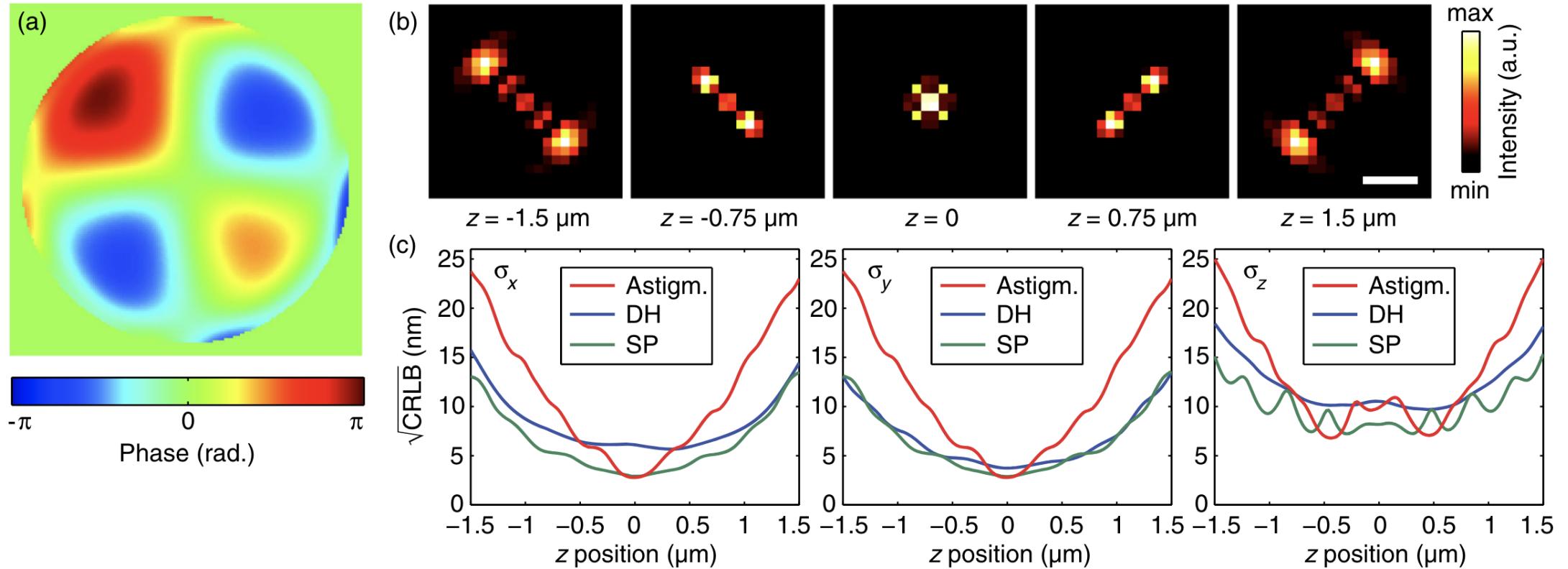


Depth is easy to estimate with DH optical systems

Designing a Phase Mask for a Conventional Camera

- Have a point source at location $\mathbf{x} = [x, y, z]^t$
- Observe $I = \text{Poisson}(h_\phi(\mathbf{x}))$, where the PSF h is function of the phase mask ϕ
- Construct the Fisher Information matrix associated with estimating \mathbf{x}
 - Error of maximum likelihood estimator of \mathbf{x} is bounded by reciprocal of Fisher Information
- Design an optimal PSF by maximizing Fisher Information wrt ϕ

Designing a Phase Mask for a Conventional Camera

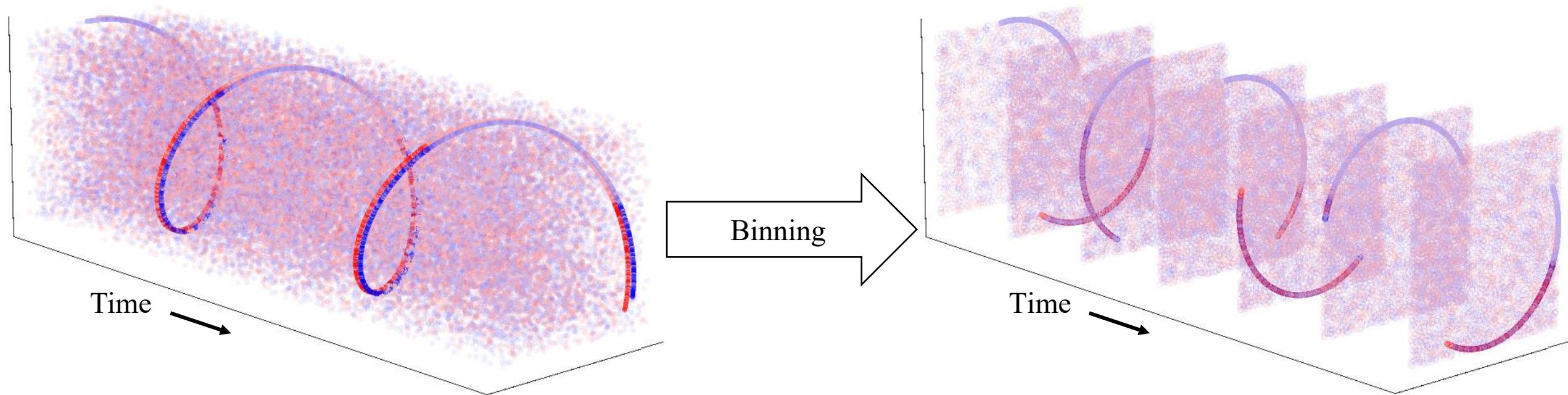


Can we extend this approach to event cameras?

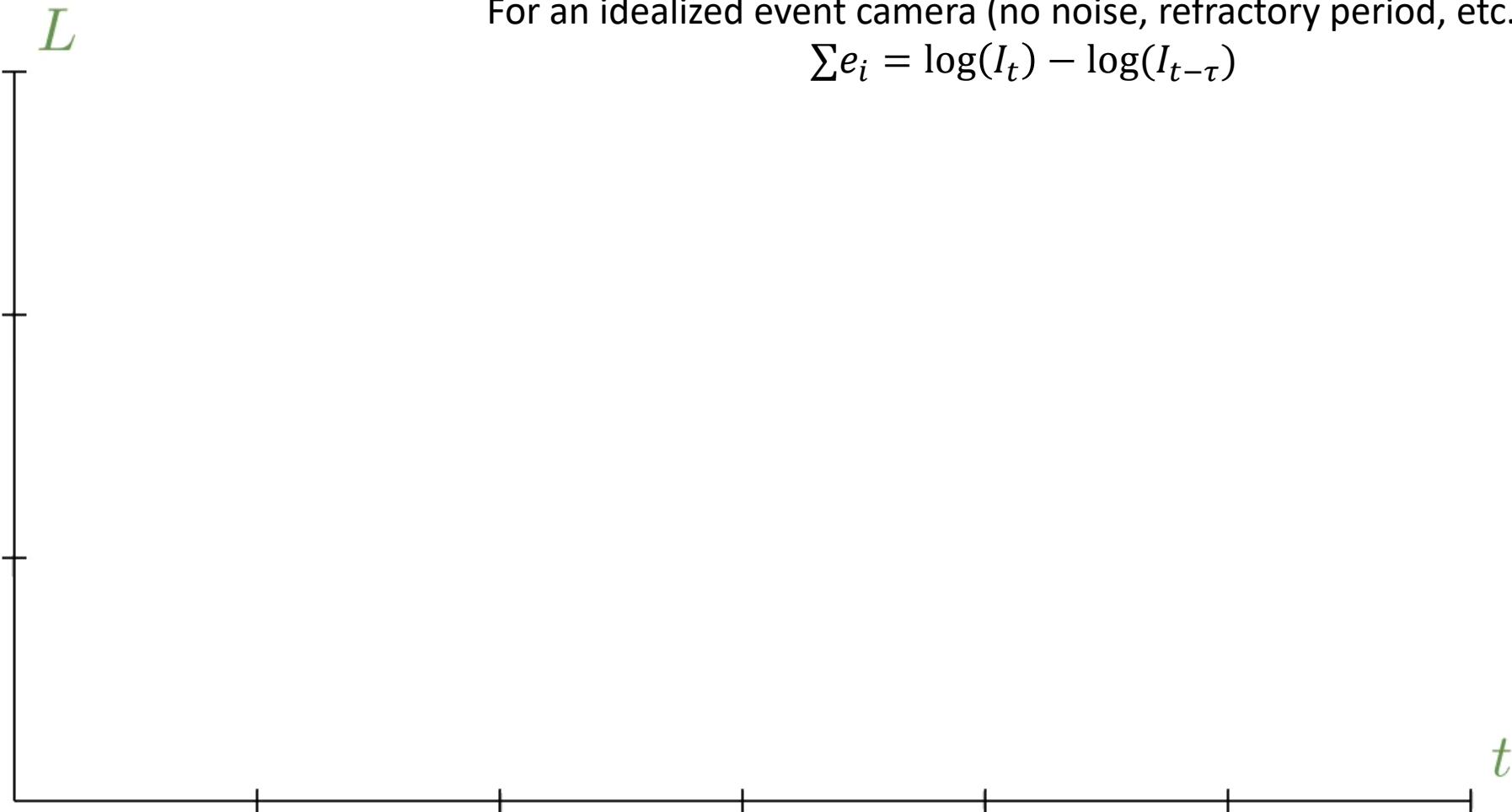
Designing a Phase Mask for an Event Camera

- Have a point source at location $\mathbf{x} = [x, y, z]^t$ moving with velocity $\Delta\mathbf{x} = [\Delta x, \Delta y, \Delta z]^t$
- Observe $I \neq \text{Poisson}(h_\phi(\mathbf{x}))$, where the PSF h is function of the phase mask ϕ
- Construct the Fisher Information matrix associated with estimating \mathbf{x}
 - Error of maximum likelihood estimator of \mathbf{x} is bounded by reciprocal of Fisher Information
- Design an optimal PSF by maximizing Fisher Information wrt ϕ

Binning Events

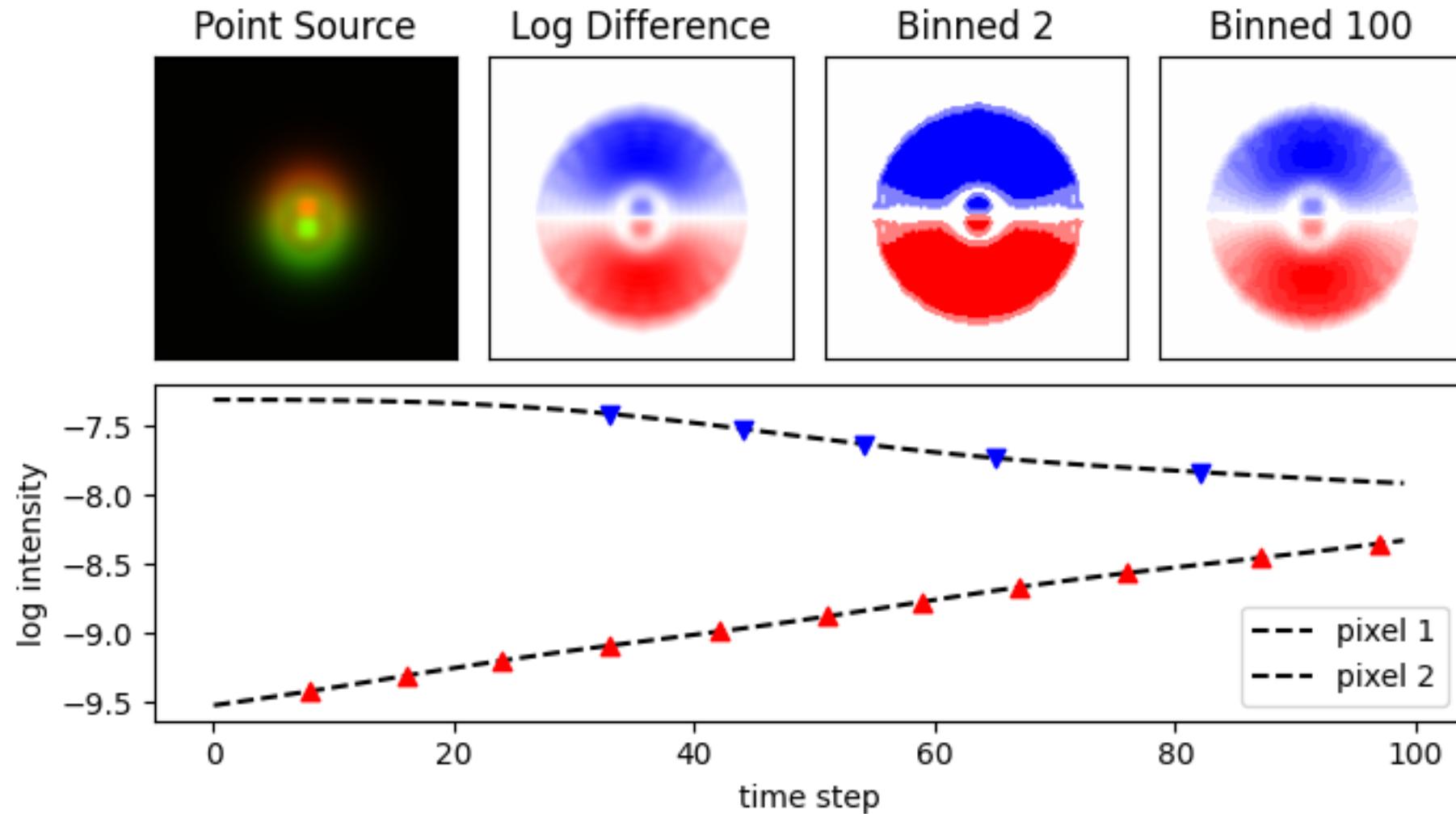


Binning Events



Binning Events

Binned events = Log difference between frames



Theory: Stationary Flashing Point Source



$$M = \log(I_t) - \log(I_{t-\tau})$$

$$= \log(I_t)$$

$$e^M = I_t \sim \text{Poisson}(\lambda = PSF)$$

Key Finding: For blinking fluorescent molecules,
the Fisher PSF is already optimal!

Theory: Generalization



$$M = \log(I_t) - \log(I_{t-\tau})$$

$$= \log\left(\frac{I_t}{I_{t-\tau}}\right)$$

$$\begin{aligned} e^M &= \frac{I_t}{I_{t-\tau}} \sim \frac{\text{Poisson}(\lambda_t)}{\text{Poisson}(\lambda_{t-\tau})} \\ &\sim \mathcal{N}\left(\frac{\lambda_t}{\lambda_{t-\tau}}, \frac{\lambda_t}{\lambda_{t-\tau}^2} + \frac{\lambda_t^2}{\lambda_{t-\tau}^3}\right) \end{aligned}$$

$$\mathcal{I}(\theta) = \sum_n^N \frac{\mathcal{D}^T \mathcal{D}}{2(\mu + \nu)^2} \odot \begin{bmatrix} a & a & a & b & b & b \\ a & a & a & b & b & b \\ a & a & a & b & b & b \\ b & b & b & c & c & c \\ b & b & b & c & c & c \\ b & b & b & c & c & c \end{bmatrix}$$

Challenge #1: Highly non-convex wrt lens parameters

Challenge #2: Depends on particle position and motion

$$[\mu x/\nu \quad \mu y/\nu \quad \mu z/\nu \quad x/\nu \quad y/\nu \quad z/\nu]$$

$$a = 2\mu^2\nu + 4\mu^2 + 2\mu\nu^2 + 12\mu\nu + 9\nu^2$$

$$b = - (2\mu^2\nu + 2\mu^2 + 2\mu\nu^2 + 7\mu\nu + 6\nu^2)$$

$$c = 2\mu^2\nu + \mu^2 + 2\mu\nu^2 + 4\mu\nu + 4\nu^2$$

$$\mathcal{I}(\theta) = \sum_n^N \frac{\mathcal{D}^T \mathcal{D}}{2(\mu + \nu)^2} \odot \begin{bmatrix} a & a & a & b & b & b \\ a & a & a & b & b & b \\ a & a & a & b & b & b \\ b & b & b & c & c & c \\ b & b & b & c & c & c \\ b & b & b & c & c & c \end{bmatrix}$$

Challenge #1: Highly non-convex wrt lens parameters

Solution #1: Regularize with INRs

Challenge #2: Depends on particle position and motion

Solution #2: Monte Carlo averaging

$$[r^{\mu x}/\mu \quad r^{\mu y}/\mu \quad r^{\mu z}/\mu \quad r^{\nu x}/\nu \quad r^{\nu y}/\nu \quad r^{\nu z}/\nu]$$

$$a = 2\mu^2\nu + 4\mu^2 + 2\mu\nu^2 + 12\mu\nu + 9\nu^2$$

$$b = -(2\mu^2\nu + 2\mu^2 + 2\mu\nu^2 + 7\mu\nu + 6\nu^2)$$

$$c = 2\mu^2\nu + \mu^2 + 2\mu\nu^2 + 4\mu\nu + 4\nu^2$$

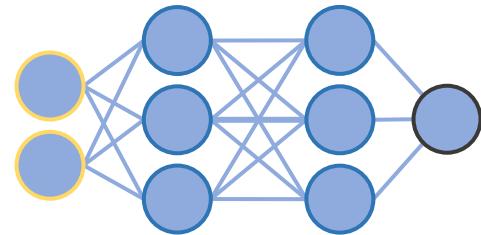
Implicit Neural Representations

Form *functional* representations of images (or phase masks)

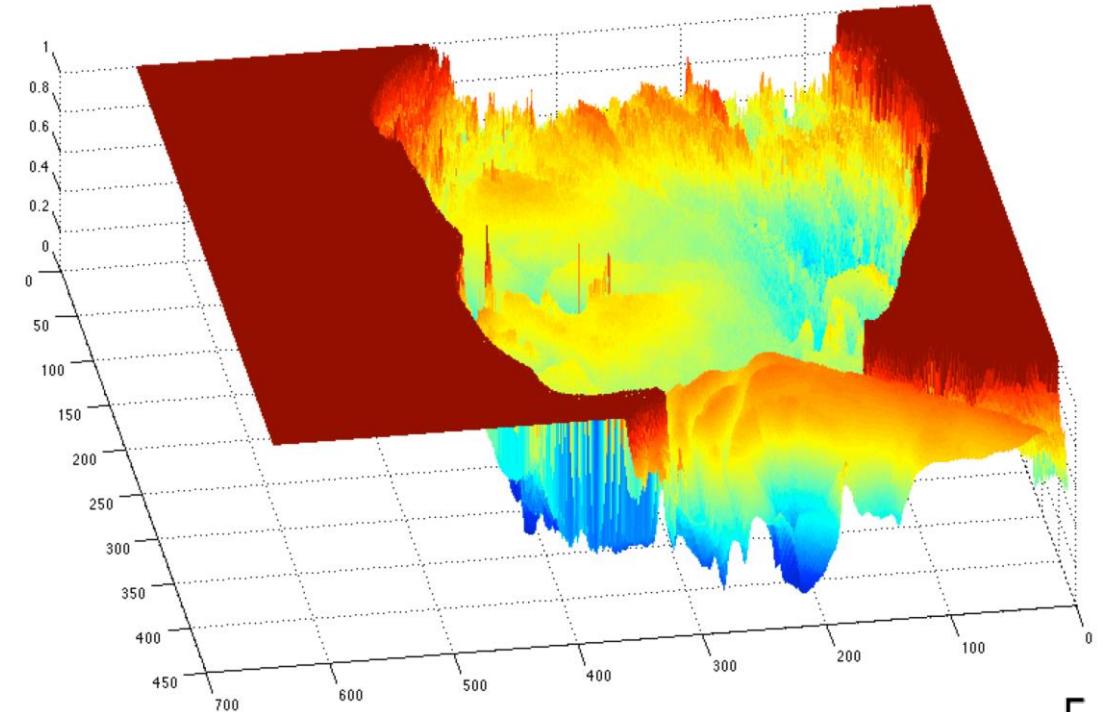
E.g., a (grayscale) image is a 2D function



grayscale image



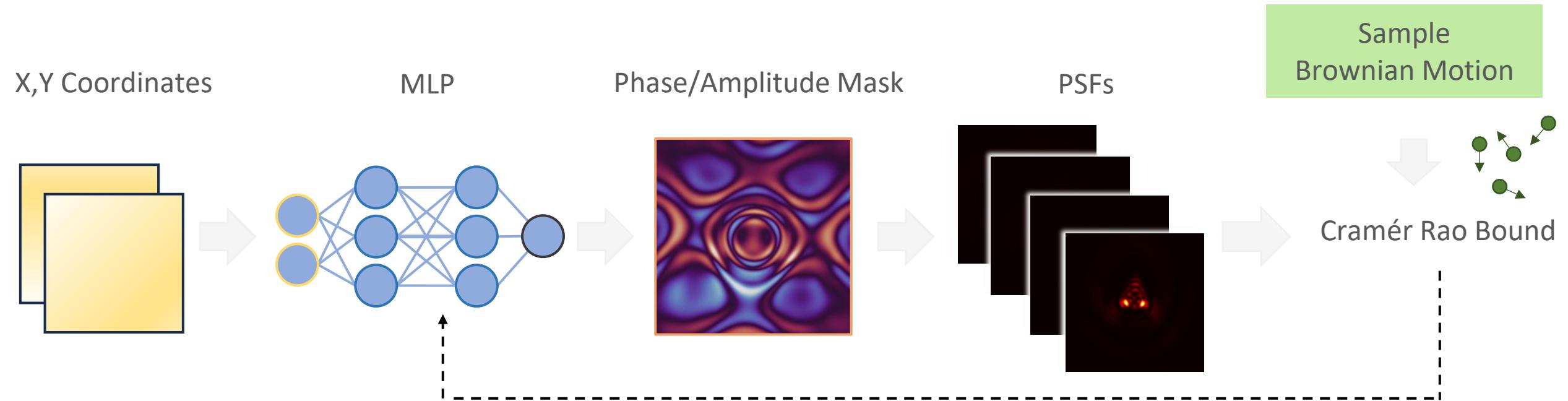
$$f(\mathbf{x})$$



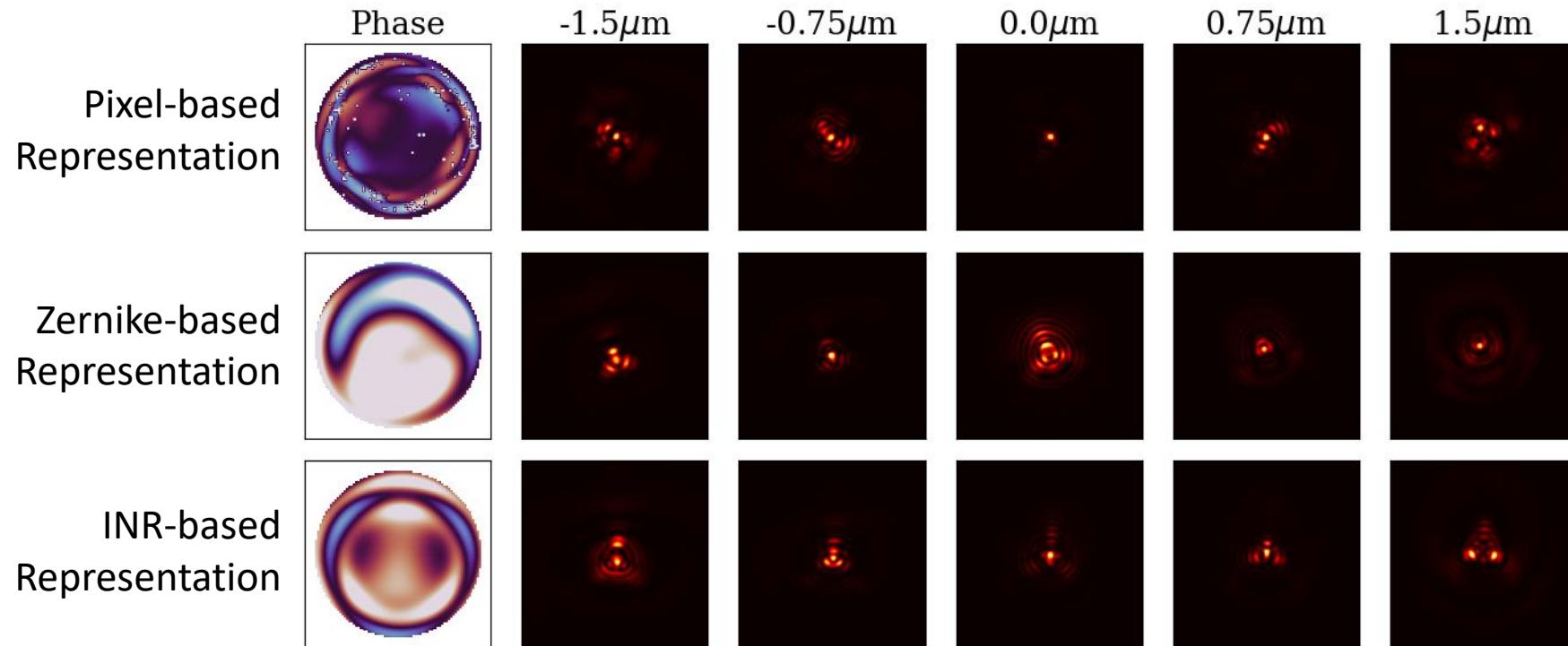
domain $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$

[Image Credit: Kris Kitani]

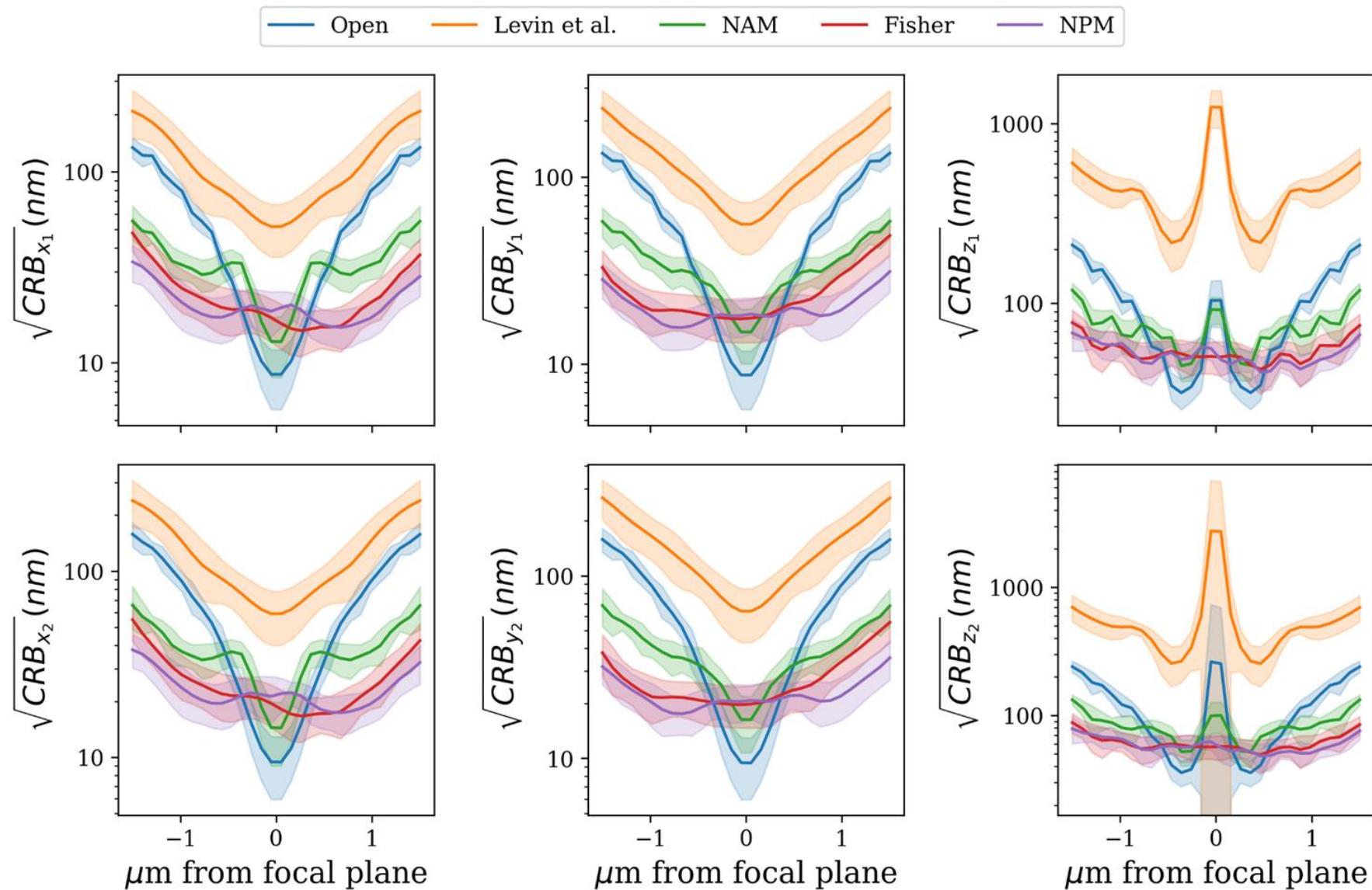
Optical Design with Implicit Neural Representations and MC Sampling



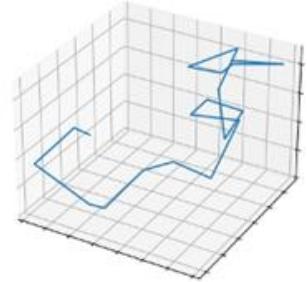
Optical Design: Learned Masks



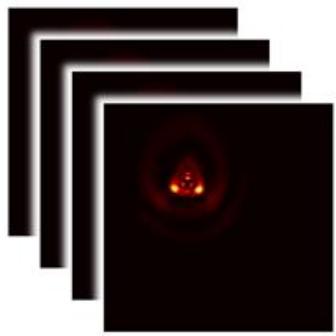
Theoretical Results



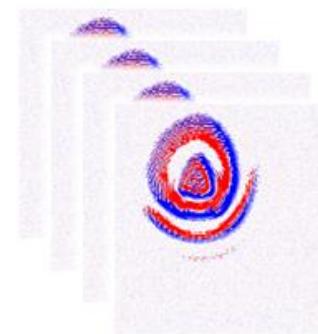
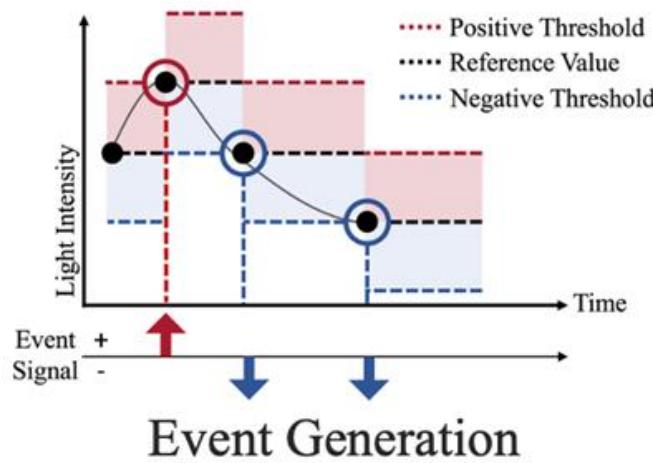
3D Tracking: Simulation Training



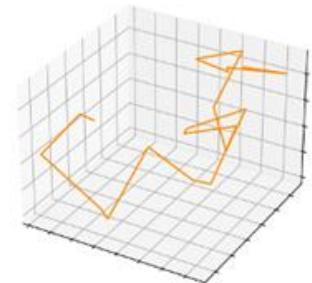
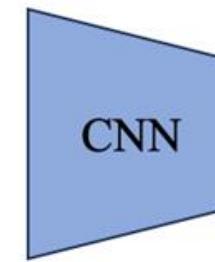
Ground Truth
Trajectory



Coded Frames

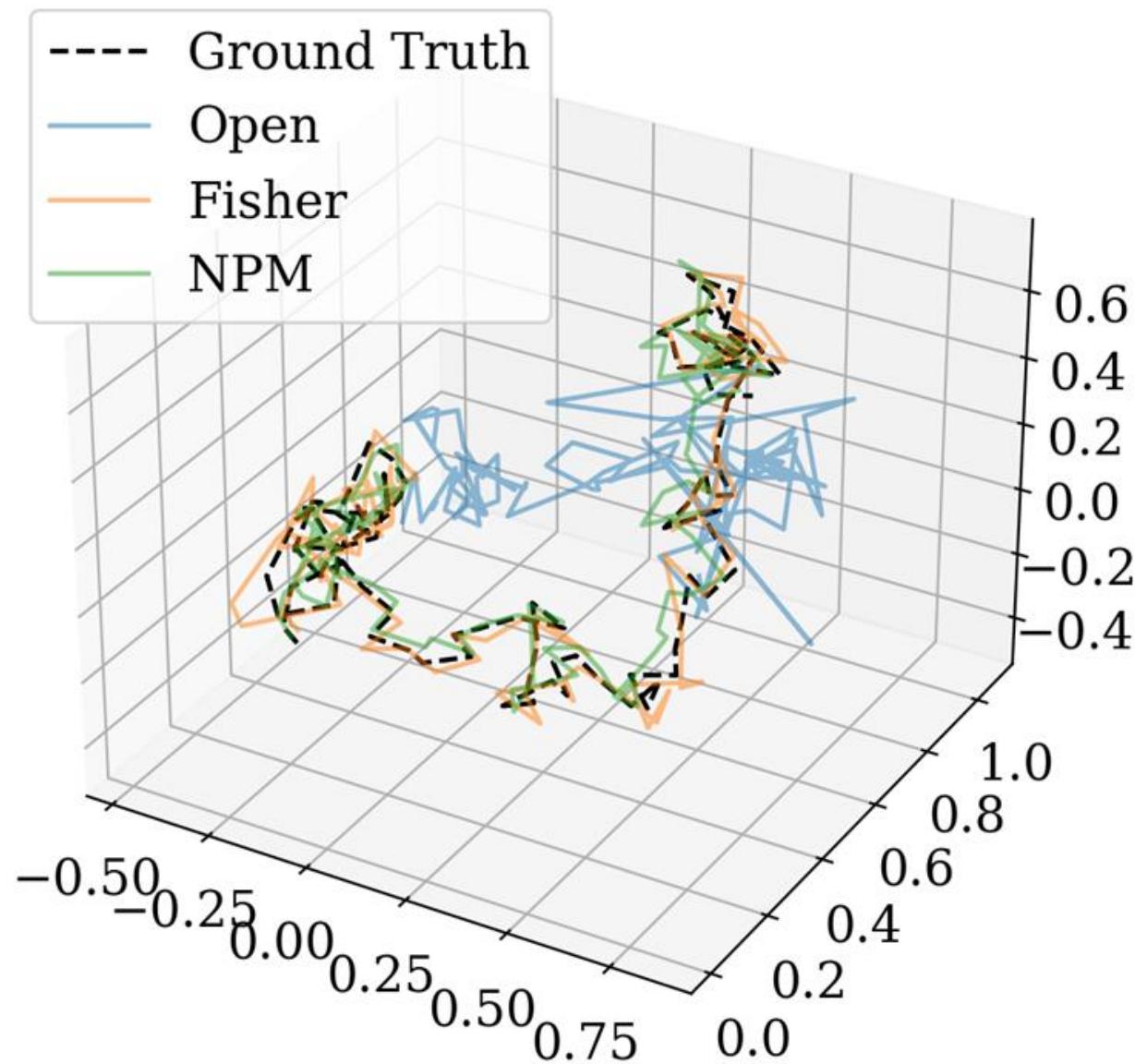


Coded Events



Recovered 3D
Trajectory

3D Tracking: Results



Lab Prototype: Setup

Prototype



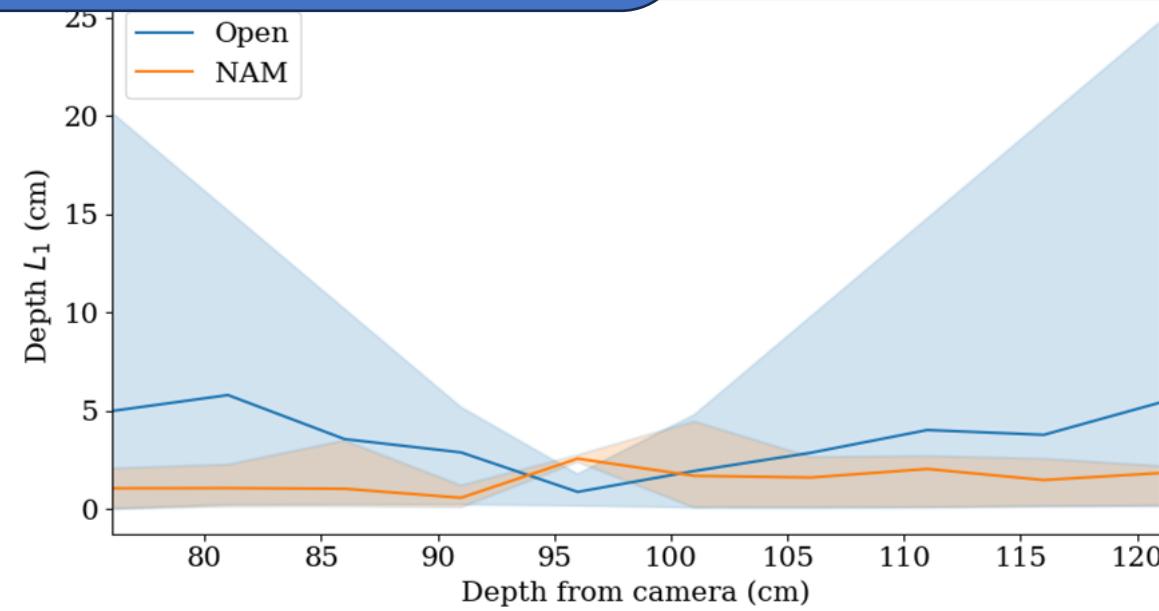
Regular Event

Coded Event

Next Steps: More sophisticated optics



Optical Setup



Todays Talk:

Part 1
Optical Encoder

Sensor
Bottleneck

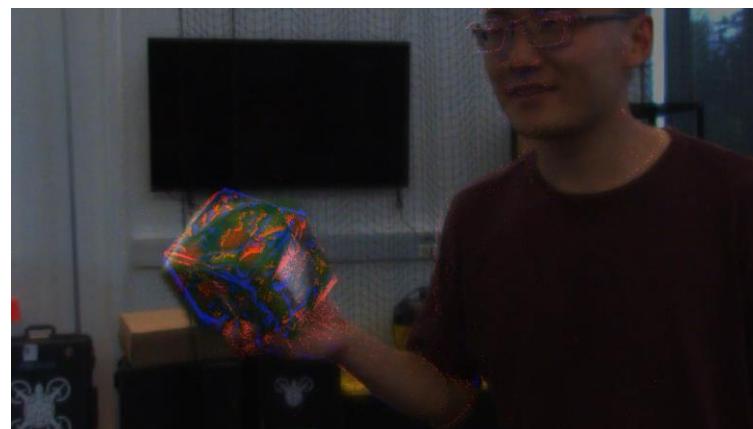
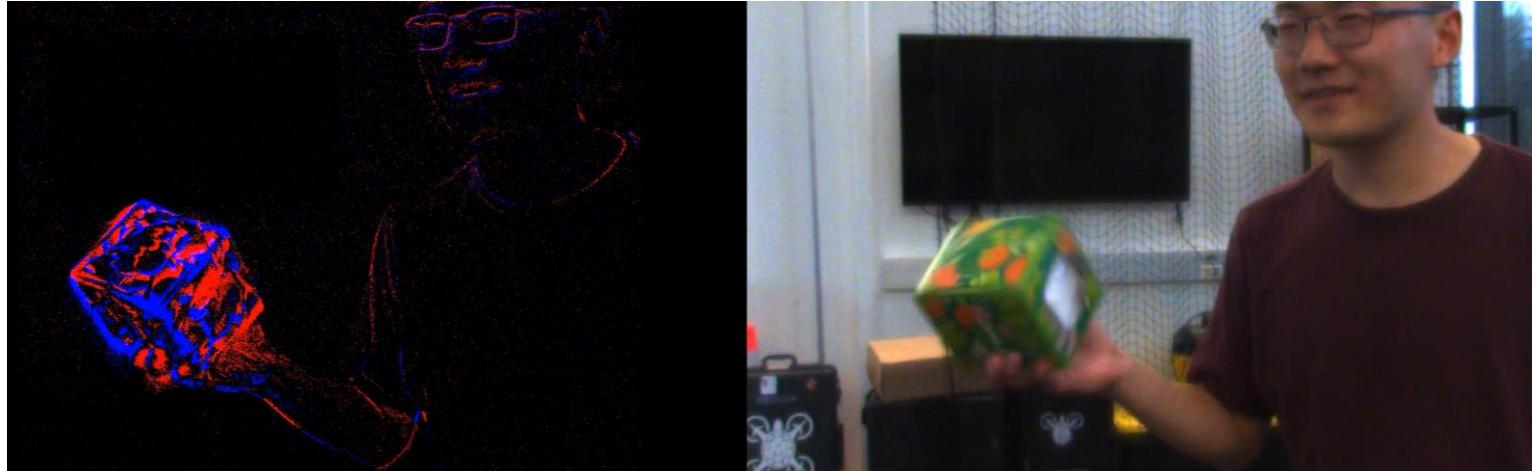
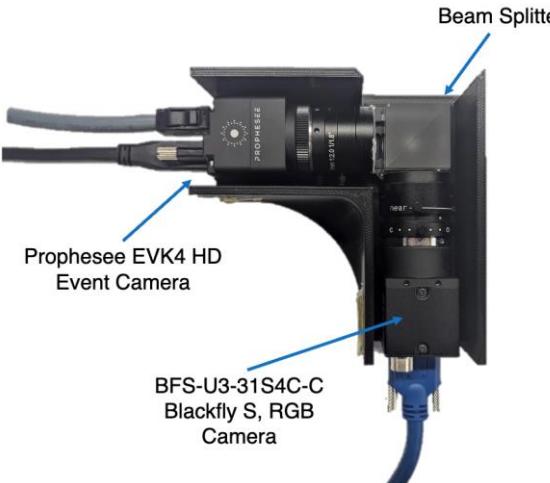
Part 2
Digital Decoder

Jingxi Chen



Event-Guided Video Frame Interpolation

- Large motion between frames makes rgb-only video frame interpolation ill-posed
- Event-based Video Frame Interpolation (EVFI) addresses this challenge by using sparse, high-temporal-resolution event measurements as motion guidance.



Related Work

Time Lens: Event-based Video Frame Interpolation

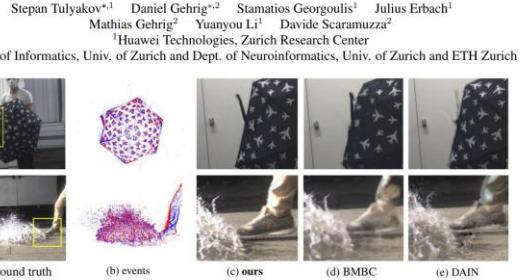


Figure 1: Qualitative results comparing our proposed method, Time Lens, with DAIN [3] and BMBC [28]. Our method can interpolate frames in highly-dynamic scenes, such as while spinning an umbrella (top row) and bursting a balloon (bottom row). It does this by combining events (b) and frames (a).

Abstract

State-of-the-art frame interpolation methods generate intermediate frames by inferring object motions in the image from consecutive key-frames. In the absence of additional information, first-order approximations, i.e. optical flow, must be used, but this choice restricts the types of motions that can be modeled, leading to errors

Time Lens++: Event-based Frame Interpolation with Parametric Non-linear Flow and Multi-scale Fusion

Stepan Tulyakov¹ Daniel Gehrig^{*2} Stamatis Georgoulis¹ Julius Erbach¹

Mathias Gehrig² Yuanyou Li¹ Davide Scaramuzza²
Huawei Technologies, Zurich Research Center

²Dept. of Informatics, Univ. of Zurich and Dept. of Neuroinformatics, Univ. of Zurich and ETH Zurich

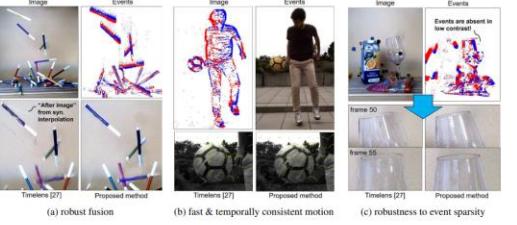


Figure 1: Comparison to state-of-the-art event- and image-based video interpolation method Time Lens [3]. Our method makes a series of key innovations to address the limitations of current approaches. First, it uses feature-level multi-scale fusion which is robust to artifacts in the fused images (a). Second, it computes continuous flow, parameterized by splines, which have inherent temporal consistency (b, bottom right vs. left) and can be efficiently sampled, thereby significantly reducing computation for multi-frame interpolation (b). Finally, it combines images and events to generate flow, even where few events are triggered, thereby mitigating artifacts as in (c).

Abstract

Recently, video frame interpolation using a combination of frame- and event-based methods has surpassed traditional frame-based methods both in terms of performance and memory efficiency. However, current methods still suffer from (i) brittle image-level fusion of complementary interpolation results, that fails in the presence of artifacts in the fused image, (ii) potentially temporally inconsistent and inefficient motion estimation procedures, that run for every inserted frame and (iii) low contrast regions that do not trigger events, thus cause events-only motion esti-

Multimedia Material

Unifying Motion Deblurring and Frame Interpolation with Events

Xiang Zhang, Lei Yu[†]
Wuhan University, Wuhan, China.
{xiangz, ly-wd}@whu.edu.cn

Abstract

Slow shutter speed and long exposure time of frame-based cameras often cause visual blur and loss of inter-frame information, degenerating the overall quality of captured videos. To this end, we present a unified framework of event-based motion deblurring and frame interpolation for blurry video enhancement, where the extremely low latency of events is leveraged to alleviate motion blur and facilitate intermediate frame prediction. Specifically, the mapping relation between blurry frames and sharp latent images is first presented by a novel learnable attention module, and a fusion network is then proposed to refine the coarse motion fields utilizing the information from consecutive blurry inputs and the concurrent events. By exploring the mutual constraints among blurry frames, latent images, and event streams, we further propose a self-supervised learning framework to enable network training with real-world blurry videos and events. Extensive experiments demonstrate that our method compares favorably against the state-of-the-art approaches and achieves remarkable performance on both synthetic and real-world datasets. Codes are available at <https://github.com/XiangZ-0/EVDI>.

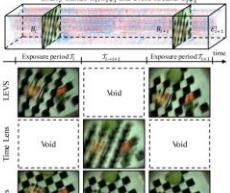


Figure 1: Illustrative examples of video deblurring and interpolation via the state-of-the-art deblurring approach LEVS [11], interpolation approach Time Lens [30] and our EVDI method.

because of motion ambiguities and the erasure of intensity textures [11]. Besides current frame-based interpolation

Event-based Video Frame Interpolation with Cross-Modal Asymmetric Bidirectional Motion Fields

Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, Kuk-Jin Yoon
Korea Advanced Institute of Science and Technology
{intelpro,yujeong,jhg0001,kjyoon}@kaist.ac.kr

Abstract

Video Frame Interpolation (VFI) aims to generate intermediate video frames between consecutive input frames. Since the event cameras are bio-inspired sensors that only encode brightness changes with a micro-second temporal resolution, several works utilized the event camera to enhance the performance of VFI. However, existing methods estimate bidirectional inter-frame motion fields with only events or approximations, which can not consider the complex motion in real-world scenarios. In this paper, we propose a novel event-based VFI framework with cross-modal asymmetric bidirectional motion field estimation. In detail, our EIP-BIOPNet utilizes each valuable characteristic of the event and images for direct estimation of inter-frame motion fields with cross-modal motion fields. Moreover, we develop an interactive attention-based frame synthesis network to efficiently leverage the complementary warping-based and synthesis-based features. Finally, we build a large-scale event-based VFI dataset, ERF-X170FPS, with a high frame rate, extreme motion, and dynamic texture to overcome the limitations of previous event-based VFI datasets. Extensive experimental results validate that our method shows significant performance improvement over the state-of-the-art VFI methods on various datasets. Our project pages are available at: <https://github.com/intelpro/CBNet>



Figure 1: Qualitative comparison on the warped frame of inter-frame motion fields. (b) and (c) estimate symmetrical inter-frame motion fields. (d) and (e) estimate asymmetric motion fields using only images and events, respectively. (f) Ours shows the best results using cross-modal asymmetric bidirectional motion fields.

motion-based VFI methods [3, 4, 8, 13, 22, 29, 30] are proposed thanks to the recent advance in motion estimation algorithms [13, 14, 16, 23, 34, 41]. For the inter-frame motion field estimation, the previous works [3, 12, 29] estimate the optical flows between consecutive frames and approximate intermediate motion fields [12, 29, 49] using linear [12, 29] or quadratic [49] approximation assumptions. These methods often estimate the inaccurate inter-frame motion fields when the motions between frames are vast or non-linear, adversely affecting the VFI performance.

More paired training data, bigger and more expressive models, better performance

Two Issues

More paired event + rgb data → more \$\$\$

Larger models → more overfitting to specific cameras and interpolation rates

In Domain:



Out of Domain:



Can we benefit from massive datasets and highly expressive models *without* having to generate or train these datasets/models ourselves?

Our Contribution: Bring in the Big Guns!

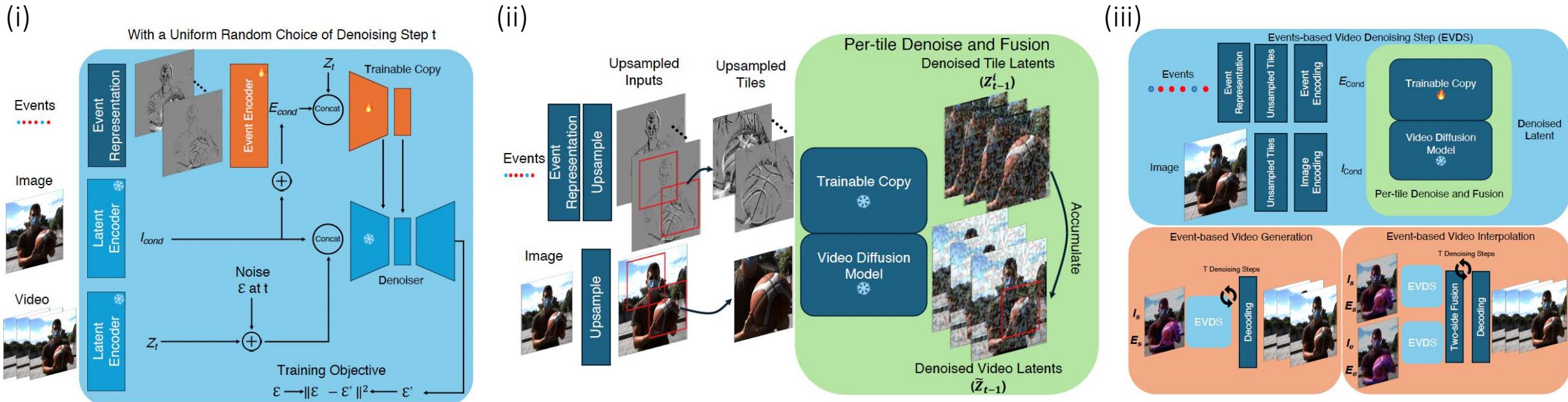
Pretrained Video Diffusion Models

1. Internet-scale Data → Strong data prior (Generalization)
2. Video Diffusion → Denoising a video at once, temporal consistency



Proposed Approach:

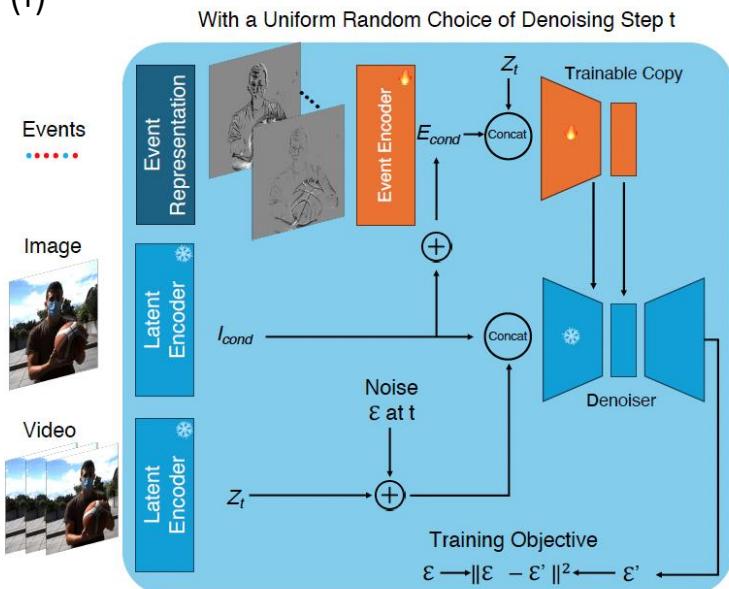
- (i) Control a pretrained diffusion model with event guidance
- (ii) Preprocess to preserve spatial and temporal resolution
- (iii) Use video *generation* network to perform video *interpolation*



Proposed Approach:

- (i) Control a pretrained diffusion model with event guidance
- (ii) Preprocess to preserve spatial and temporal resolution
- (iii) Use video *generation* network to perform video *interpolation*

(i)



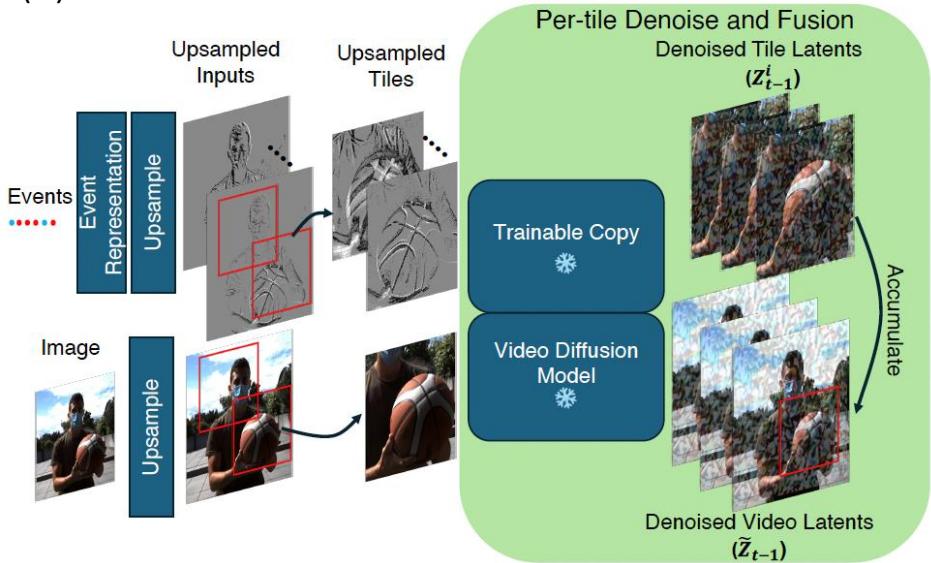
ControlNet-style approach adapts a small event-to-latent encoder while preserving the original video diffusion models weights.

Can learn to control diffusion model with only a limited amount of training data, without the risk of forgetting the original video priors

Proposed Approach:

- (i) Control a pretrained diffusion model with event guidance
- (ii) Preprocess to preserve spatial and temporal resolution
- (iii) Use video *generation* network to perform video *interpolation*

(ii)



The diffusion models encoding process is inherently lossy

Can preserve fine details by upsampling inputs before encoding

Without upsampling



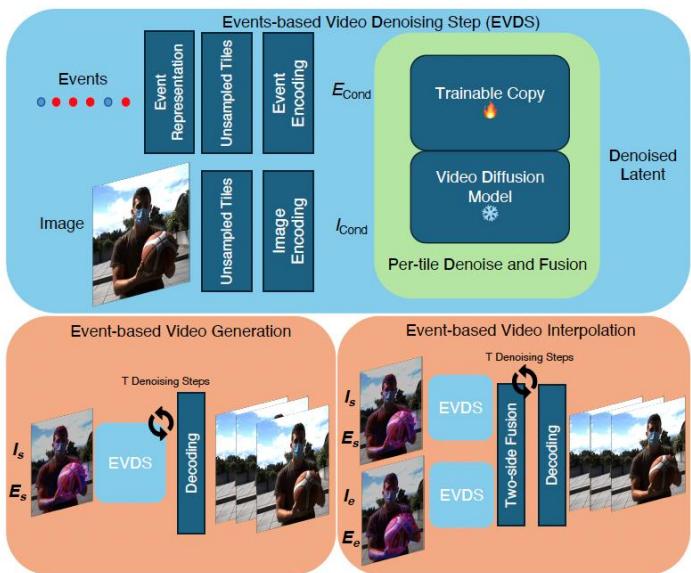
With upsampling



Proposed Approach:

- (i) Control a pretrained diffusion model with event guidance
- (ii) Preprocess to preserve spatial and temporal resolution
- (iii) Use video *generation* network to perform video *interpolation*

(iii)



Video diffusion model performs generation, not interpolation

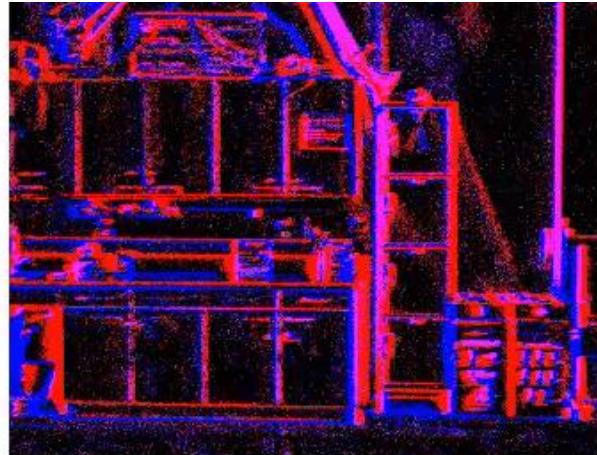
To interpolate, run video model forward (from first frame) and in reverse (from last frame). Blend latents.



Results: Generalizes to Extreme Interpolation without Fine Tuning



Input Frames



Input Events



RIFE



Time Reversal



Reference



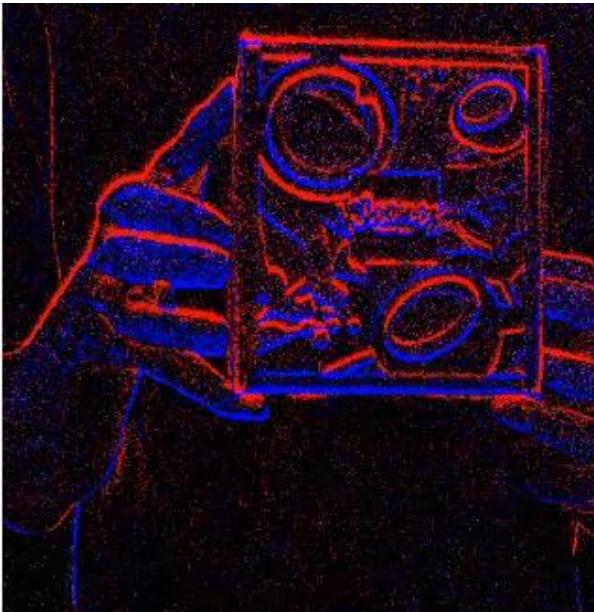
CBMNet-Large



Ours



Input Frames



Input Events



RIFE



Time Reversal



Reference

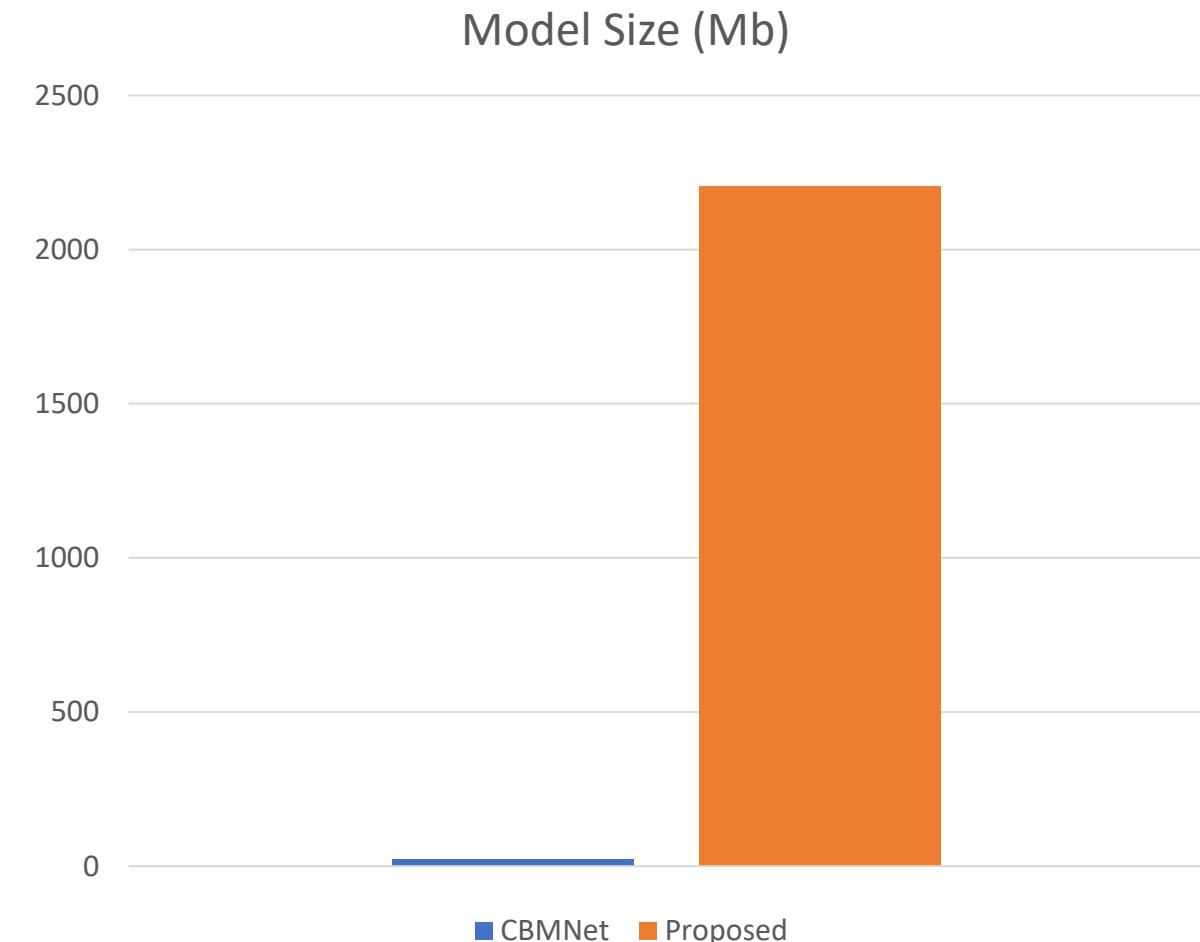
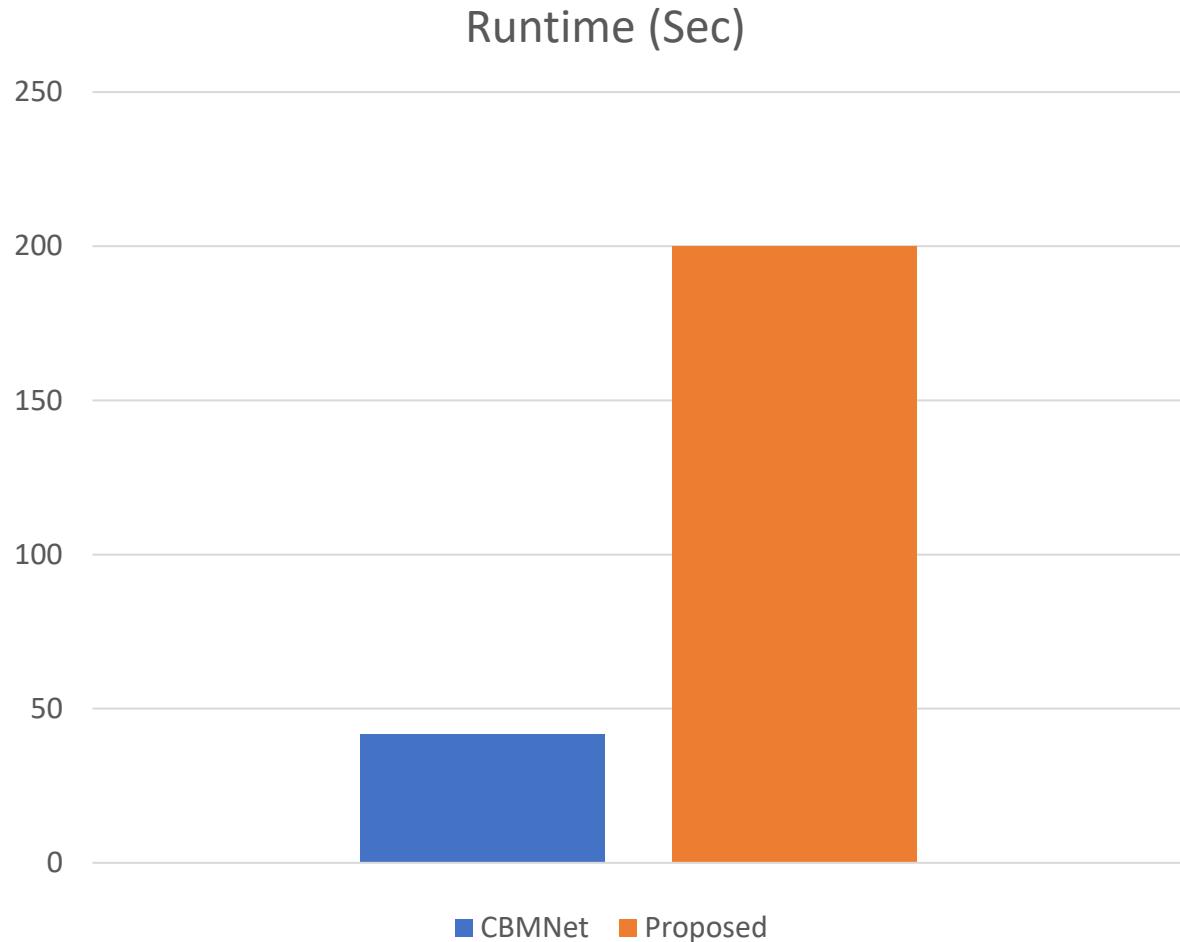


CBMNet-Large



Ours

Downsides?



Todays Talk:

Part 1
Optical Encoder

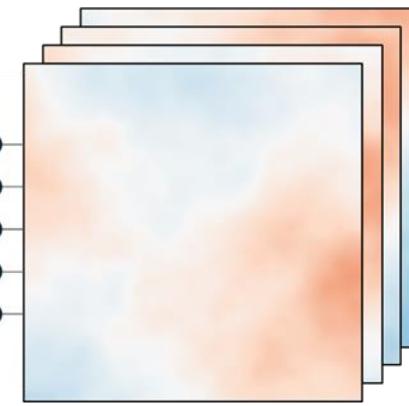
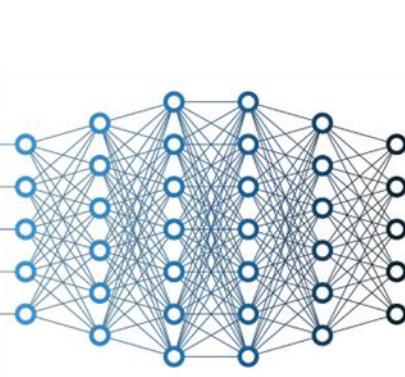
Sensor
Bottleneck

Part 2
Digital Decoder

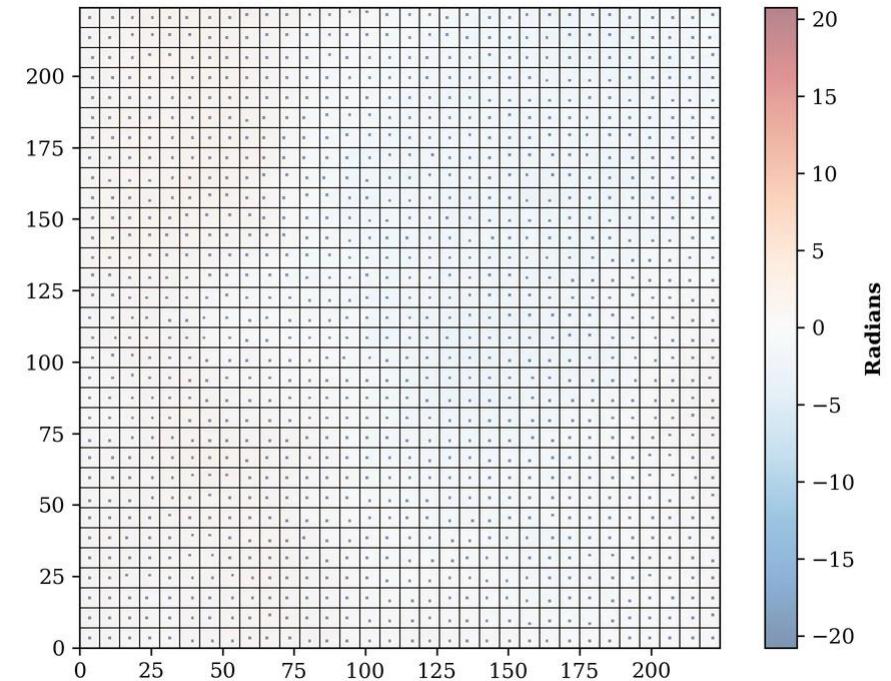
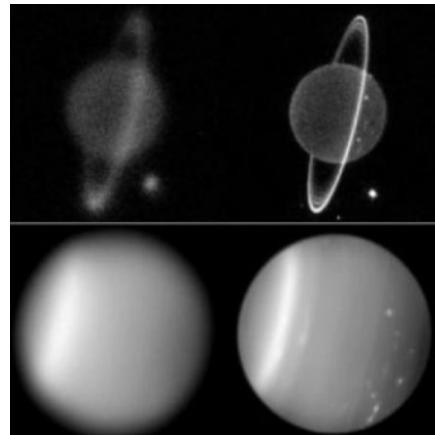
High-speed predictive wavefront sensing with event-cameras



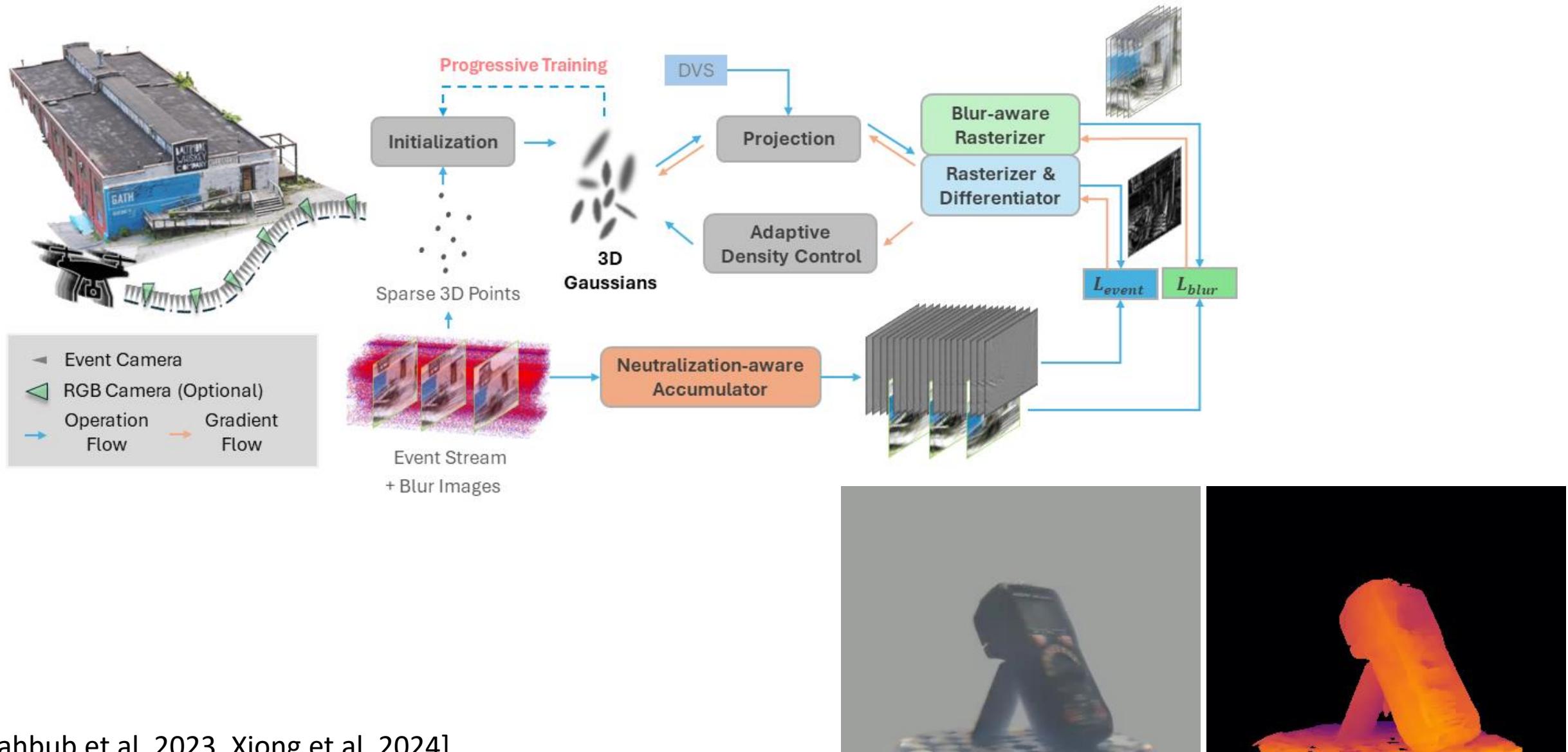
**Event-based
Wavefront Sensing**

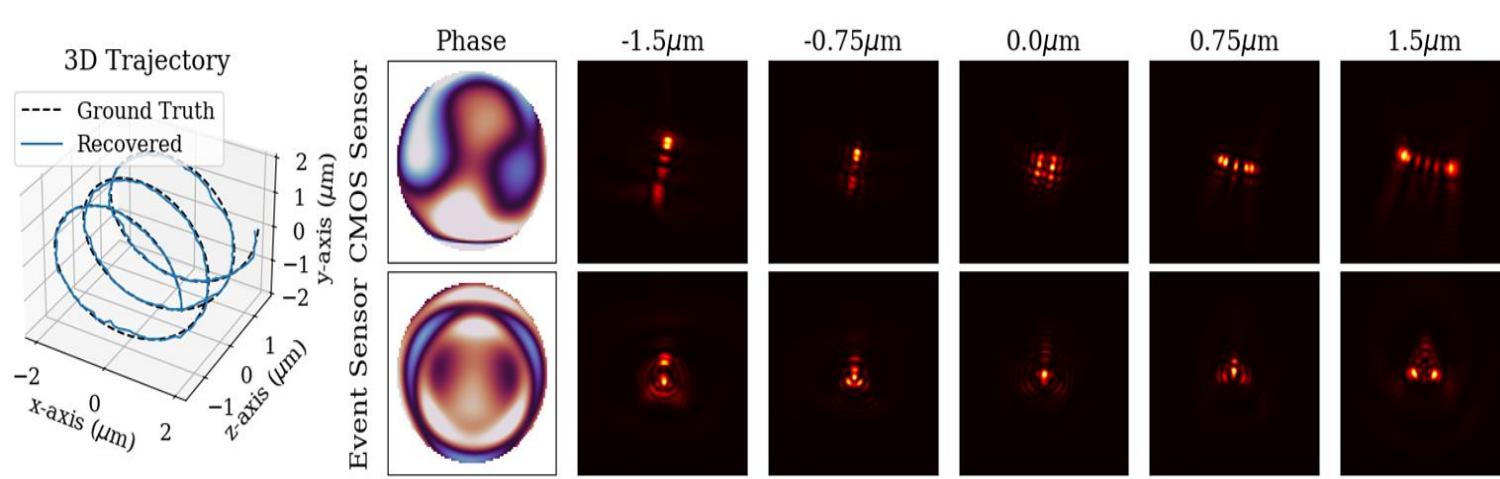


**Deep Predictive
Reconstruction**



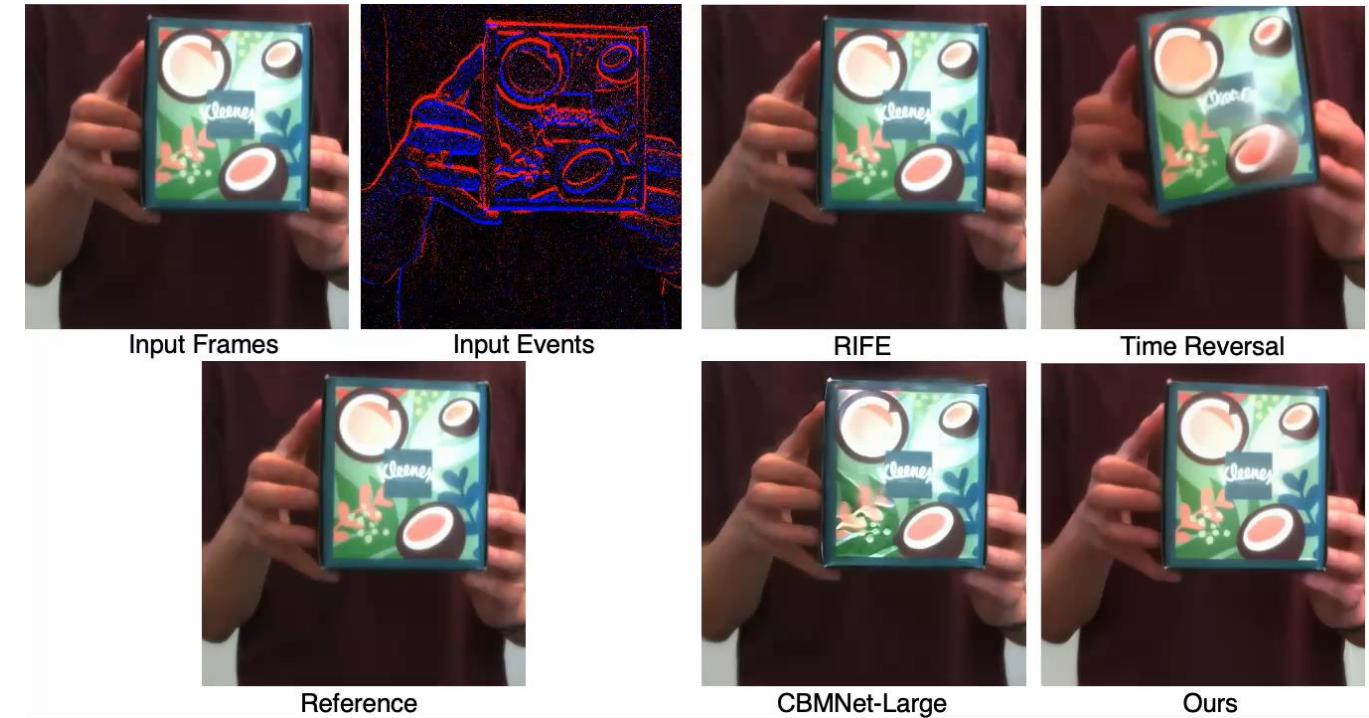
Dense 3D Reconstruction with Inverse Differentiable Rendering





Event cameras + coded optics

Event cameras + generative models



References:

CodedEvents: Optimal Point-Spread-Function Engineering for 3D-Tracking with Event Cameras *CVPR 2024*

Repurposing pre-trained video diffusion models for event-based video interpolation *CVPR 2025. Sat. morning poster session*

Acknowledgements



Sachin Shah



Matthew
Chan



Haoming
Cai



Jingxi Chen



Sakshum
Kulshrestha



Chahat
Deep Singh



Brandon Feng



Tianfu
Wang



Levi
Burner



Dehao
Yuan

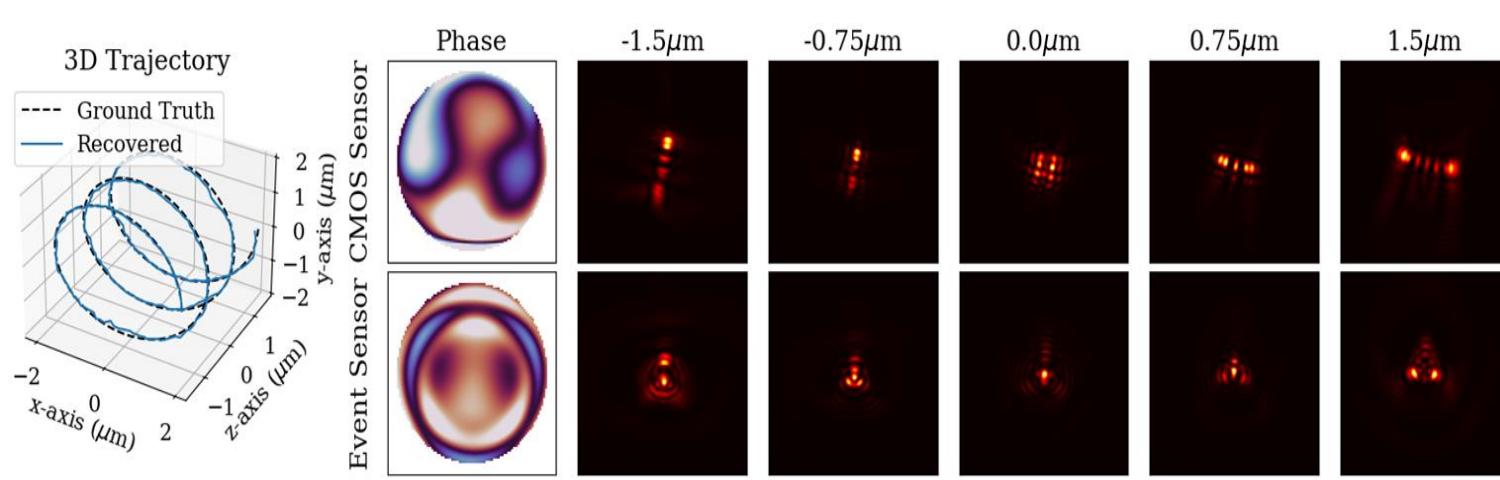


Cornelia
Fermuller



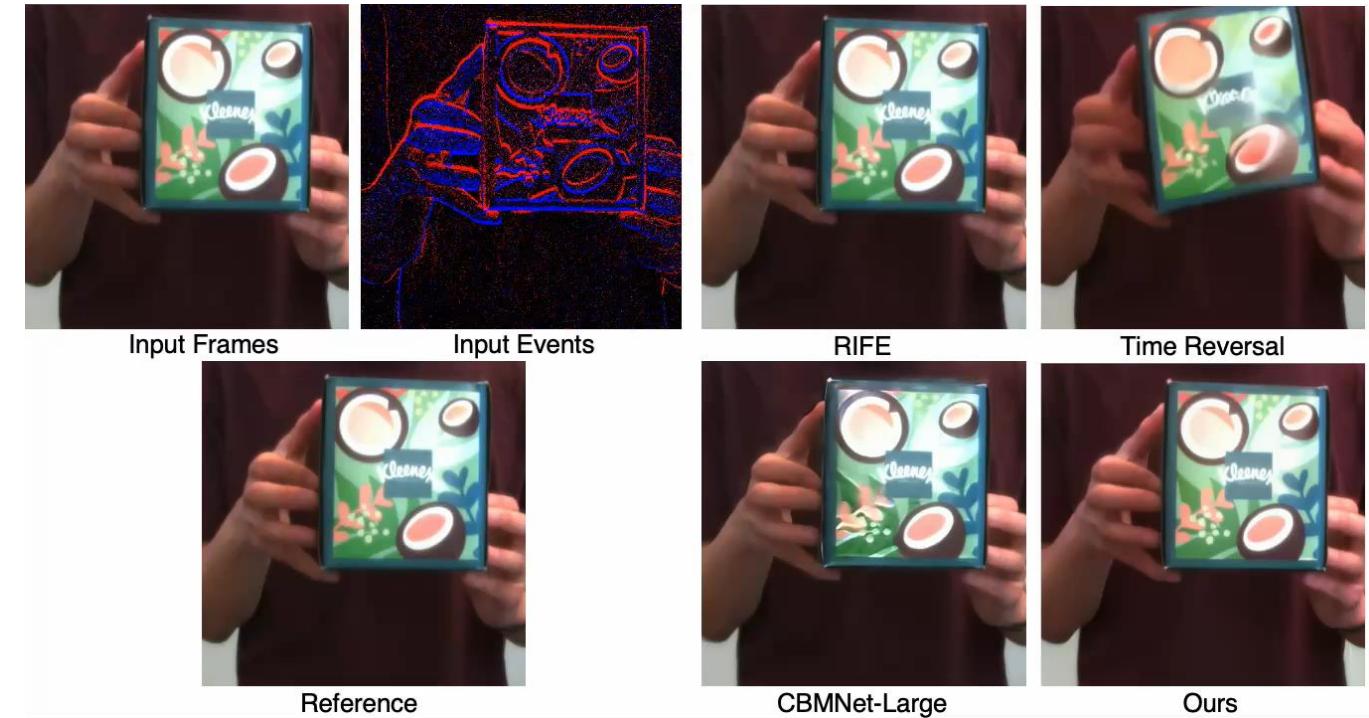
Yiannis
Aloimonos





Event cameras + coded optics

Event cameras + generative models



References:

CodedEvents: Optimal Point-Spread-Function Engineering for 3D-Tracking with Event Cameras *CVPR 2024*

Repurposing pre-trained video diffusion models for event-based video interpolation *CVPR 2025. Sat. morning poster session*