

## Event Guided Depth Sensing

Manasi Muglikar<sup>1</sup>

Diederik Paul Moeys<sup>2</sup>

Davide Scaramuzza<sup>1</sup>

### Abstract

Active depth sensors like structured light, lidar, and time-of-flight systems sample the depth of the entire scene uniformly at a fixed scan rate. This leads to limited spatio-temporal resolution where redundant static information is over-sampled and precious motion information might be under-sampled. In this paper, we present an efficient bio-inspired event-camera-driven depth estimation algorithm. In our approach, we dynamically illuminate areas of interest densely, depending on the scene activity detected by the event camera, and sparsely illuminate areas in the field of view with no motion. The depth estimation is achieved by an event-based structured light system consisting of a laser point projector coupled with a second event-based sensor tuned to detect the reflection of the laser from the scene. We show the feasibility of our approach in a simulated autonomous driving scenario and real indoor sequences using our prototype. We show that, in natural scenes like autonomous driving and indoor environments, moving edges correspond to less than 10% of the scene on average. Thus our setup requires the sensor to scan only 10% of the scene, which could lead to almost 90% less power consumption by the illumination source. While we present the evaluation and proof-of-concept for an event-based structured-light system, the ideas presented here are applicable for a wide range of depth sensing modalities like LIDAR, time-of-flight, and standard stereo. Video is available at <https://youtu.be/Rvv9IQLYjCQ>.

### 1. Introduction

High-resolution and high-speed depth estimation is desirable in many applications such as robotics, AR/VR, and 3D modeling. Active depth sensors such as structured light, light detection and ranging (LiDAR), and time-of-flight

<sup>1</sup>Dept. Informatics, University of Zurich and Dept. Neuroinformatics, University of Zurich and ETH Zurich, Switzerland. <sup>2</sup>Advanced Sensors and Modelling Group, SONY R&D Center Europe, SL1. This research was supported by SONY R&D Center Europe and the National Centre of Competence in Research (NCCR) Robotics, through the Swiss National Science Foundation.

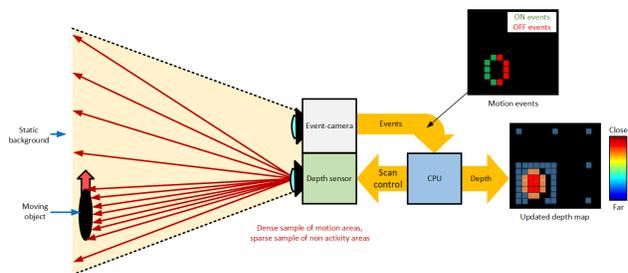


Figure 1: System overview. Events are triggered by moving objects in the scene or by relative motion. These events guide the depth sensor to scan with higher spatial density the affected motion areas of the field of view, resulting in denser depth regions. In the rest of the field of view, the depth sensor scans at a lower spatial density, resulting in sparser depth regions.

(ToF) scan the entire scene uniformly at fixed scan rates. This leads to limited spatio-temporal resolution where redundant static information is over-sampled and precious motion information might be under-sampled. We take inspiration from human perception, which involves foveating: areas of interest are scanned with higher spatial resolution while the other regions are scanned with lower spatial resolution. The area of interest for robotics applications is often characterized by high-contrast features or dynamic objects in the scene; therefore, identifying these areas with low-latency in challenging environments is essential. We propose to use an event camera to identify these areas of interest as events naturally correspond to moving objects (assuming brightness constancy) and, therefore, do not require further processing to segment the area of interest.

Our proposed system is illustrated in Fig. 1. The scene consists of a static background with an object moving in front of it. The motion of the scene triggers events in the event camera corresponding to the edges of the moving object. These events, represented as an event-frame, are used to guide the depth sensor to scan with higher spatial density the affected motion areas, leading to denser depth samples for the moving object. The rest of the field of view is scanned sparsely, leading to sparser depth samples for the background. In scenarios where the setup also undergoes ego-motion (e.g., driving), events would correspond

Method	Depth sensor	Modality	Frame rate	Adaptive
Intel RealSense	SL (stereo)	I	90	✗
Azure Kinect	ToF	I	30	✗
Programmable light curtains [5]	SL	I	5.6	✓
Pittaluga et al. [6]	ToF	I	24	✓
Brandli et. al. [1]	SL	E	-	✗
Matsuda et. al. [2]	SL	E	60	✗
Martel et. al. [3]	SL(stereo)	E	-	✗
Muglikar et. al. [4]	SL	E	60	✗
<b>Ours</b>	SL	E	60	✓

Table 1: Summary of active depth sensors. The modality indicates the sensor used in the setup for depth estimation ( $I$  is for standard camera images and  $E$  stands for event camera).

to edges of the scene, triggered by moving objects or relative motion. Thus, the depth samples would have higher resolution in the neighbourhood of scene edges and sparse resolution in constant-intensity regions of the scene.

As a depth-sensing modality, in this work we consider event-based structured light (SL). Event-based SL systems provide high-speed depth estimation compared to their frame-based counterparts [1–4] due to the high temporal resolution of event cameras. However, their accuracy also drops at higher spatial resolutions due to noise in event data. While we present the evaluation and proof-of-concept for an event-based SL system, the ideas presented here are applicable for a wide range of depth sensing modalities like LIDAR, ToF, and standard stereo. Note that we do not want to replace full-frame depth sensors, but rather provide a complementary depth sensor that can scan dynamic objects with low-latency. We summarize our **contributions** below:

- A novel formulation of event-guided foveating depth sensing to guide a depth sensor to scan moving objects with higher spatial resolution and low-latency.
- Real-world experimental evaluation on dataset recorded with a prototype.
- Neural-network based depth completion from event-guided sparse depth samples.

The paper is structured as follows: In Section 2.1, we give an overview of event-based structured light systems. Section 3 presents the problems with high resolution and high frequency event-based SL systems and non-idealities of event cameras. Section 5 details our proposed solution for event-guided depth estimation. In Section 5.1 we provide details for depth completion from sparse event-guided depth. Finally, we evaluate our method in simulation and real world in Section 6 and conclude in Section 7.

## 2. Related works

Prior work on active depth sensors is summarized in the Table 1. The first half focuses on frame-based active depth sensors (modality  $I$ ) and the second half focuses on

event-based active depth sensors (modality  $E$ ). We first discuss literature on event-based active depth in Section 2.1 followed by a review of adaptive sampling depth sensors (frame-based SL or ToF) in Section 2.2. Lastly, we summarize some of the other sensing modalities that make use of foveation for different applications in Section 2.3.

### 2.1. Event-based active depth

Since event cameras are novel sensors (commercially available since 2008), the literature on event-based structured light systems is not vast. Table 1 summarizes some of the event-based SL systems. The earliest work [1] combined a DVS with a pulsed line laser to reconstruct a terrain. The pulsed laser line was projected at a fixed angle with respect to the DVS while the terrain moved beneath it. In [3], they combined a laser light source with a *stereo* setup consisting of two DAVIS240 cameras [7]. The laser illuminated the scene and the synchronized event cameras recorded the events generated by the reflection from the scene. Hence the light source was used to generate stereo point correspondences, which were then triangulated (back-projected) to obtain a 3D reconstruction. The SL system MC3D [2] comprised of a laser point projector and a DVS. The laser scanner moved in a raster pattern across the scene. The DVS captured the reflection of the light from the scene as events. MC3D then converted temporal information of events at each pixel into disparity. It exploited the redundancy suppression and high temporal resolution of the DVS, also showing appealing results in dynamic scenes compared to frame-based counterparts. Previous systems demonstrated the capabilities of event camera for SL with a low-resolution event sensor. Recently, [4] used a high resolution event camera for SL and discuss the noises present in event timestamps (jitter and BurstAER effects). They proposed a depth estimation algorithm to improve the robustness of event-based SL up to a certain extent. However, as higher resolution event cameras are emerging, the accuracy of event-based SL systems does not scale with the resolution as noise in the event timestamps increases significantly due to the high event rate.

### 2.2. Adaptive depth

One of the earliest works used programmable light curtains to detect objects at fixed shapes around the robot or car [5]. Recently, [6] showed a proof of concept prototype with camera-based foveating LIDAR showing the advantage of the motion-based adaptive sampling. However, they were limited by the hardware and camera to  $24Hz$  depth acquisition time.

### 2.3. Adaptive foveating sensors

The concept of attention-driven sensing is inspired by biology. It follows the idea that the human eye more effi-

ciently concentrates its processing in areas of high interest (high contrast or motion) while under-sampling redundant information (static part of the background or low contrast features). In the retina, the attention area is focused in the fovea, which is densely populated with photoreceptors and retinal ganglion cells [8]. This approach has found interest in computer vision for various algorithms including 3D SLAM [9, 10]. To avoid rendering entire frames, foveation has also been used in AR/VR headsets [11]

### 3. Event-based SL

In this section, we introduce the basics of event-camera-projector setup. We discuss the sources of noise (Section 3.1.1) and evaluate the theoretical performance of this system at high resolutions and projector scanning frequencies (Section 3.1.2).

#### 3.1. Basics of event-based SL

We consider the problem of depth estimation using a laser point projector and event camera. The scene is illuminated by the laser beam inside the projector, moving in a raster scanning fashion. This light reflected from the scene generates events<sup>1</sup> in the event camera. Assuming the travel time for light from the projector to the camera is small, we can map the event pixel to the corresponding projector pixel using the event timestamp. Depth is estimated by triangulating the pixel from the event-camera plane and the projector plane [4]. Depth, thus, depends on accurate event timestamps. We now take a look at the non-idealities of events that lead to noise in the timestamps.

##### 3.1.1 Event camera imprecision sources

Available event cameras have a minimum possible temporal resolution of 1  $\mu$ s. However, in practice, latency, transistor noise and variable readout delays introduce noise in the data, which decrease the precision of the event timestamp. Such non-idealities are design and manufacturer specific. *Latency*: defined as the time it takes for an event to be registered since the moment logarithmic change in intensity exceeds the threshold, it can currently range on average from a few microseconds to hundreds of milliseconds, depending on bias settings, manufacturing process and illumination level. *Transistor noise*: defined as the random transistor noise of the circuits, it also depends on settings and illumination. This noise randomly changes the measured signal, leading to threshold-comparison jitter [12]. *Other*

<sup>1</sup>An event camera generates an event  $e_k = (\mathbf{x}_k, t_k, p_k)$  at time  $t_k$  when the logarithmic brightness at the pixel  $\mathbf{x}_k = (x_k, y_k)^\top$  increases or decreases by predefined threshold  $C$ :  $L(\mathbf{x}_k, t_k) - L(\mathbf{x}_k, t_k - \Delta t_k) = p_k C$ , where  $p_k \in \{-1, +1\}$  is the sign (polarity) of the brightness change, and  $\Delta t_k$  is the time since the last event at the same pixel location.

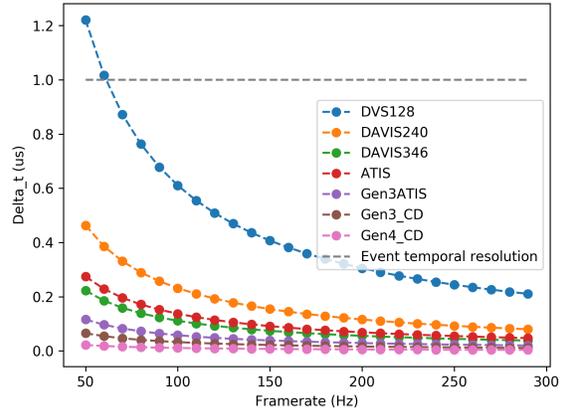


Figure 2: Theoretical effect of projector frame rate and resolution on time between consecutive events. We observe that with an increase in the framerate from 50 Hz to 100 Hz, for the DVS128 (lowest resolution) sensor, the time between two consecutive events drops below the event temporal threshold of 1  $\mu$ s

*non-idealities*: encompassing parasitic photocurrents and junction leakages, these effects bias event generation to one specific polarity [13]. *Read-out architecture*: arbitrated architectures, which preserve or partially preserve the order of the pixels’ firing, lead to significant queuing delays before the timestamping operation. This is particularly noticeable when the number of active pixels (and resolution) scale up. Scanning readouts, on the other hand, limit the possible delays by sacrificing event timing resolution. *Jitter*: defined as the random variation that appears in timestamps, it depends on all of the aforementioned factors, all of which increase the unpredictability and imprecision of the event timing.

##### 3.1.2 Performance at high resolution and scanning frequencies

Accuracy of event-based SL system depends on the accuracy of event timestamps. Therefore, for consecutive events triggered by the laser beam to be uniquely distinguished, the timestamps should differ by at least 1  $\mu$ s, which is the event camera timestamp resolution. The time it takes for laser beam of the projector to move to the next pixel,  $\Delta t$ , is inversely proportional to the scanning frequency  $f$  and spatial resolution  $W \times H$  pixels:

$$\Delta t = \frac{1}{f \times W \times H} \quad (1)$$

We define this approach as dense sampling since it illuminates every pixel. We consider commercially available event cameras covering a wide range of resolutions namely:

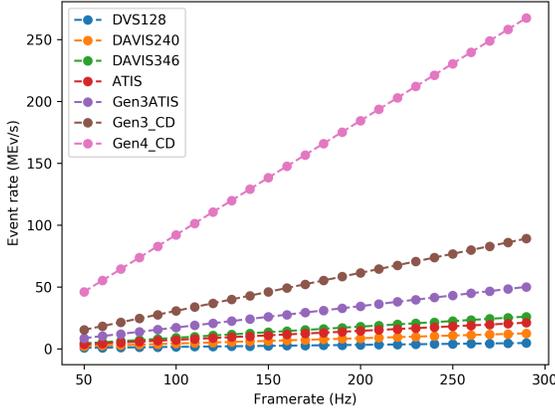


Figure 3: Theoretical effect of projector frame rate and resolution on event-rate of an event-based structured light system.

DVS128 (resolution  $128 \times 128$ ), DAVIS240 (resolution  $240 \times 180$ ), DAVIS346 (resolution  $346 \times 260$ ), Prophesee ATIS (resolution  $302 \times 240$ ), Prophesee Gen3.CD (resolution  $640 \times 480$ ), Prophesee Gen3ATIS (resolution  $480 \times 360$ ), Prophesee Gen4.CD (resolution  $1280 \times 720$ ), and compute the theoretical  $\Delta t$  for each sensor while varying the projector framerate from 50 Hz to 290 Hz. Projector framerate here refers to the scanning frequency of the laser point projector. As we see in Fig. 2,  $\Delta t$  is lower than  $1 \mu\text{s}$  for almost every sensor resolution for scanning frequency higher than 60 Hz. Since  $\Delta t$  depends on the scanning frequency and sensor resolution, there is no simple way to ensure  $\Delta t$  is higher than event camera temporal resolution.

Similarly, another disadvantage of dense sampling arises from the high-event rates. In event-based SL, the event-rate (measured in MEv/s) is directly proportional to the scanning frequency and spatial resolution. Theoretical event-rates for multiple projector scanning frequencies and event camera resolutions are shown in Fig. 3. For lower resolution cameras such as DVS128 [14], the increase in the event-rate with respect to the scanning frequency is not significant. Whereas for a higher resolution camera such as Prophesee Gen4 [15], the event rate increases from 44 MEv/s (scanning frequency of 50 Hz) to 265 MEv/s (scanning frequency of 290 Hz). Since event rate plays a significant role in the event timestamp error [16], the errors in depth will be significant in high-resolution and high-speed event-based structured light systems. At a high event rate, of 10 MEv/s, the timestamp error is in the order of  $10 \mu\text{s}$  which can go up to  $200 \mu\text{s}$  for event rates of 265 MEv/s. Sparse sampling mitigates some of these errors by reducing the event rate.

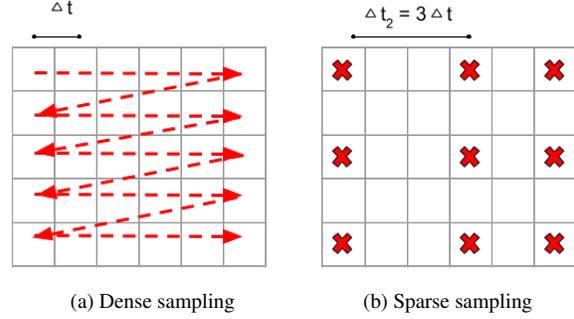


Figure 4: Visualization of dense sampling versus sparse sampling. The  $\Delta t$  between consecutive events for dense sampling is very small. On the other hand,  $\Delta t$  for sparse sampling is 3 times higher than dense in this example.

## 4. Sparse sampling

Sparse sampling increases the  $\Delta t$  between consecutive events by scanning a fraction of the pixels as shown in Fig. 4. The  $\Delta t$  is now proportional to the number of pixels skipped between two illuminated pixels (defined as  $N$ ) as shown in Equation. 2. This allows us to control the timestamp errors due to small  $\Delta t$  up to a certain extent.

$$\Delta t = \frac{N}{f \times W \times H} \quad (2)$$

Sparse sampling, however, does not consider the scene geometry. This can lead to undersampling of areas of fast moving objects and oversampling planar areas, leading to redundant information. Conversely, sparse sampling is not always desirable in areas of interest (usually of object motion), as seen in Fig. 5a. In an autonomous driving scenario, for example, if a fast moving vehicle or pedestrian moves in front of a car, it is useful to have more depth information about this. For robotics applications, dynamic objects in the scene are important for obstacle avoidance or local planning of trajectories. These scenarios make it desirable to have a higher resolution in these areas of interest at a low-latency.

## 5. Event-guided depth sensing

Sparse sampling has the disadvantage of under sampling areas of interest and over sampling planar areas. We therefore propose to use an event-based sampling for structured light depth with event cameras. As events naturally encode the relative motion of the scene with high temporal resolution, they are ideal for this task. Our approach, is shown in Fig. 1. The motion of the scene triggers events in the event camera corresponding to the edges of the moving object. Assuming brightness constancy, events observed by the event camera correspond to edges moving either due

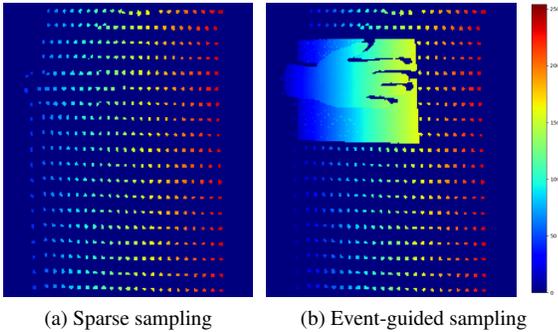


Figure 5: Example event time map with sparse sampling and event-guided sampling.

to ego-motion or object motion. The distinction between these two types of events (*i.e.* ego-motion events and object motion events) is simple in the case of a stationary camera since all the events are caused due to object motion. On the other hand, differentiating these two motion types is not simple in the case of a moving setup and is typically solved using optimization [17] or by assuming only rotational motion using IMU priors [18]. Since our application aims for low-latency depth sensing, we consider all the events (ego-motion and object motion) as equal contributors and do not distinguish between them. Future work could embed fine-tuned faster algorithms for ego motion segmentation onto ASICs or FPGAs in order to reduce latency overhead. These events, represented as an event-frame, are used to guide the depth sensor to scan with higher spatial density the affected motion areas, leading to denser depth samples for the moving object. The event camera is aligned with the depth sensor via rectification. The rest of the field of view is scanned sparsely, leading to sparser depth samples for the background. In practice, we apply a median-filter on the event-frame to reduce the influence of noise. On this event-frame, we use simple contour detection to get a bounding box around events. Fig. 5b shows the sampling strategy described here. Specifically for event-based structured light system, the depth sensor consists of a laser beam projector and a second event camera tuned to observe the events generated by the illumination source. Assuming the event camera is synchronized with the projector, events due to the projector in one scan period are represented as a time surface [19](example time surface shown in Fig. 5).

### 5.1. Depth completion with event-based sampling

While we conservatively sample the scene, a natural question to ask is "can we recover the complete depth profile from sparse depth measurements?" Several works address the problem of full scene depth estimation from sparse samples [20–23]. Recently, [24] formulated the depth com-

pletion problem with compressive sensing literature and provided exactness and stability analysis for 3D depth completion. Proposition 6 of [24] showed that exact recovery of depth from sparse set of samples is possible if the sparse sample set covers the edges and their adjacent pixels.

Previous approaches estimated dense depth from monocular events [25] or combination of events and frames [26]. In our approach, we supplement the events with depth samples generated with event-guided approach to estimate dense depth.

**Network Architecture** The general network architecture is inspired from U-Net [25, 26], which was used previously used for monocular depth estimation with events and combining events with images. Similar to [26] our architecture uses skip connections and residual block. However, unlike their approach, we do not combine events with images, but rather with event-guided depth samples.

**Depth representation:** The metric depth  $D_m$  is first converted into a normalized log depth map  $D \in [0, 1]$  facilitating large depth variation learning [25–28]. Logarithmic depth is computed similar to [26] as:

$$D = \frac{1}{\alpha} \log \frac{D_m}{D_{max}} + 1 \quad (3)$$

where,  $D_{max} = 1000$  m and  $\alpha = 5.7$ .

**Event representation:** Events are converted to fixed-size tensor to utilize existing convolutional neural network architectures. Events in a time window  $\Delta T$  are drawn into a voxel grid with spatial dimensions  $H \times W$  and  $B$  temporal bins [29–31].

$$V(x, y, t) = \sum p_i \delta(x - x_i, y - y_i) \max\{0, 1 - |t - t_i^*|\} \quad (4)$$

where  $t_i^* = \frac{B-1}{\Delta T}(t_i - t_0)$ . The number of bins  $B$  is set to 5 for all experiments [30].

## 6. Experiments

To the best of our knowledge, there is no dataset available for event-based structured light systems to evaluate the proposed method. We therefore evaluate our method by building the prototype and collecting a custom dataset. Details about the prototype are presented in Section 6.1. We also evaluate our method in simulation with EventScape dataset, a new dataset with events, intensity frames, semantic labels, and dense depth maps recorded in the CARLA simulator [25]. Since our method is agnostic to depth sensing modality (SL, ToF or LiDAR), we show the advantage of event-guided depth sensing over dense sampling. Finally, we show the result of depth completion on the simulation dataset.

### 6.1. Prototype

Our prototype for event-based structured light shown in Fig. 6 consists of a Prophesee Gen3 event camera with res-



Figure 6: Our prototype for event-based SL system consists of Prophesee Gen3 event camera (mounted on white plate) and Sony MP-CL1A laser point projector. The second event camera (mounted on red plate ) guides the laser projector to illuminate the active areas of the scene. This second event camera is next to the projector and aligned via rectification and alignment.

olution of  $640 \times 480$ . The projector is a laser point projector Sony MP-CL1A with resolution of  $1920 \times 1080$  and scanning frequency of 60 Hz. The calibration of the event camera uses [32]. We use a second Prophesee Gen3 event camera (resolution  $640 \times 480$ ) for guiding the SL depth sensor. This event camera is separated from the projector with a baseline of 4 cm. The event camera is aligned with the projector via rectification and alignment. For each sequence, alignment is performed using feature-based homography estimation by projecting a checkerboard pattern on the scene and representing the events as an event frame to extract the corner point. Note that while this alignment would work, it requires the objects to move in a front-parallel plane. To allow for alignment mismatch, we dilate the pixels in the projector plane to illuminate a larger field of view. To ensure the second event camera only observes the events and not the events due to projector illumination, we adjust the bias settings such that we filter out high frequency changes primarily caused by the projector. In particular, we reduce the bandwidth of the source follower buffer, which follows the logarithmic photoreceptor.

## 6.2. Evaluation

We evaluate the proposed approach in simulation and with a prototype. We compare the area of pixels that is required to scan with our approach in Section 6.2.1 in both simulation and real world sequences. We evaluate the performance of our approach against dense and sparse sampling techniques with our prototype in Section 6.2.2. Lastly, we show the results of our depth completion network in Section 6.3.

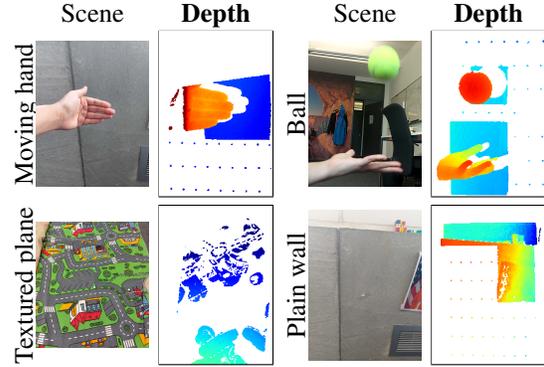


Figure 7: Reconstruction of dynamic scenes with real prototype.

### 6.2.1 Active pixels

Active pixels define the area to be scanned. For an event-guided sensing, this area is data driven. If the scene activity is present in all the pixels, this method would be equivalent to dense scanning. However, in a natural scene, the active area corresponding to moving objects is usually small compared to the sensor’s field of view (FoV). We evaluate the activity of scene on the image plane in two scenarios (*i*) stationary setup and moving objects (*ii*) autonomous driving scenario where the sensor is moving (ego-motion) along with moving objects in the scene. Since we do not differentiate between events from object motion and events due to ego-motion, we associate all the events to areas with activity.

**Simulation** For simulation, we use the EventScape dataset which demonstrates the autonomous driving scenario. Since events arise both due to ego-motion and moving objects, the active pixels depend on the scene. We calculate active pixels for over 116 test sequences. The average active pixels over all the sequences is 7.1% of the resolution (camera resolution of  $256 \times 512$ ) For the sequences that have significant motion distributed all over the scene such as  $T_{064}$ , this still amounts to 20% of the total resolution as seen in Fig. 8.

**Real Dataset** We evaluate our method on multiple dynamic scenes Fig. 7 with diverse motion scenarios to show our system’s capabilities. The percentage of active pixels in these sequences are in the range of 0.73% to 10.34%. In the stationary scenario, active pixels are minimum (0.73%) as they correspond to sparse samples. The maximum active pixels are triggered (10%) with the moving setup as this creates significant activity in the entire scene due to ego-motion. By sampling a small area of pixels, this system can lead to a low-power and high framerate depth sensor.

Method	Event rate	Recons. Error
Dense	6.8MEv/s	1.1 cm
Sparse	265kEv/s	0.76 cm
Event-guided	600kEv/s - 5MEv/s	0.89 cm

Table 2: Reconstruction of plane with dense, sparse and propose event-guided.

### 6.2.2 Reconstruction

The reconstruction results can be seen in Fig. 7. The areas without motion are sparsely sampled whereas areas with motion have dense depth. Since we post-process the events and apply a bounding box to the active area, we can also include multiple regions of interest (RoI) as shown in the Ball sequence. For the textureless wall, the motion of the setup generates events only in the areas where there are edges (*e.g.* the edge of the wall and the poster edge); therefore only those areas are scanned with high spatial density.

**Plane fitting error** We reconstruct a plane at a distance of 2 m using dense sampling, sparse sampling and event-based sampling. The event rate (measured in  $Ev/s$ ) significantly affects the event timestamp errors, as discussed in Section 3.1.1. We compare the event rates and reconstruction error for three different sampling methods: dense, sparse, and event-based while scanning a plane wall. Dense methods have a very high event rates of 6.8MEv/s, whereas the sparse sampling has a lowest event rate of 265kEv/s. The event rate for event-guided depth is data dependent and for the wall scene this varies between 600kEv/s and 5MEv/s depending on the motion. Error is computed as the plane fitting error for the reconstruction. The plane fitting error for dense sampling method is 11 mm, event-based sampling is 8.9 mm and sparse sampling is 7.6 mm. We observe that with sparse sampling, we achieve a lower reconstruction error because the event timestamps are accurate to produce accurate depth maps. One the other hand, dense sampling has the highest reconstruction error due to the noise in the event timestamps present at such high event rates. Our method strikes a balance between these two approaches and achieves a low reconstruction error and a low event-rate.

### 6.3. Depth completion

We test our depth completion network on the test sequences of EventScape dataset. Some qualitative results are shown in Fig. 8. As we see the events provide the sparse depth information to our network which learns to inpaint the depth. Since events correspond to moving edges, the sparse depth measurements also correspond to the edges and therefore the output of our depth prediction network is sharp around the edges. However in areas which are textureless and planar like  $T_{013}$  no events are generated and therefore no sparse measurements are generated leading to wrong

depth estimates in such areas. The proposed depth completion method can recover full frame depth from sparse event-guided depth input with an accuracy of 1.32 cm.

## 7. Conclusion

We present a novel bio-inspired event-guided depth estimation method. Inspired from the human ability to focus with higher resolution on areas of interest and decrease the resolution for background, we propose event-camera based depth sampling as events naturally encode relative motion of the scene. We show that in natural scenes like autonomous driving and indoor environments moving edges correspond to less than 10% of the scene. Thus our setup requires the sensor to scan only 10% of the scene. This would equate to almost 90% less power consumption by the illumination source. We demonstrate the capabilities of this system by building a prototype and qualitatively demonstrate the depth reconstruction in office scenarios.

## References

- [1] Christian Brandli, Thomas Mantel, Marco Hutter, Markus Höpflinger, Raphael Berner, Roland Siegwart, and Tobi Delbruck, "Adaptive pulsed laser line extraction for terrain reconstruction using a dynamic vision sensor," *Front. Neurosci.*, vol. 7, p. 275, 2014.
- [2] Nathan Matsuda, Oliver Cossairt, and Mohit Gupta, "MC3D: Motion contrast 3D scanning," in *IEEE Int. Conf. Comput. Photography (ICCP)*, pp. 1–10, 2015.
- [3] Julien N. P. Martel, Jonathan Müller, Jörg Conradt, and Yulia Sandamirskaya, "An active approach to solving the stereo matching problem using event-based sensors," in *IEEE Int. Symp. Circuits Syst. (ISCAS)*, pp. 1–5, 2018.
- [4] Manasi Muglikar, Guillermo Gallego, and Davide Scaramuzza, "ESL: Event-based structure light," in *IEEE International Conference on 3D Vision (3DV)*, December 2021.
- [5] Jian Wang, Joseph R. Bartels, William L. 'Red' Whittaker, Aswin C. Sankaranarayanan, and Srinivasa G. Narasimhan, "Programmable triangulation light curtains," in *Eur. Conf. Comput. Vis. (ECCV)*, 2018.
- [6] Francesco Pittaluga, Zaid Tasneem, Justin Folden, Brevin Tilton, Ayan Chakrabarti, and Sanjeev J Koppal, "Towards a mems-based adaptive lidar," *arXiv preprint arXiv:2003.09545*, 2020.
- [7] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck, "A 240x180 130dB 3μs latency global shutter spatiotemporal vision sensor," *IEEE J. Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, 2014.
- [8] Lee Ann Remington, *Clinical Anatomy and Physiology of the Visual System*. ScienceDirect, 2012.
- [9] B. Charrow, G. Kahn, S. Patil, S. Liu, P. Abbeel K. Goldberg, N. Michael, and V. Kumar, "Information theoretic planning with trajectory optimization for dense 3d mapping," *Robotics: Science and Systems*, vol. 11, 2015.
- [10] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. D. Reid, and J. J. Leonard, "Past, present, and

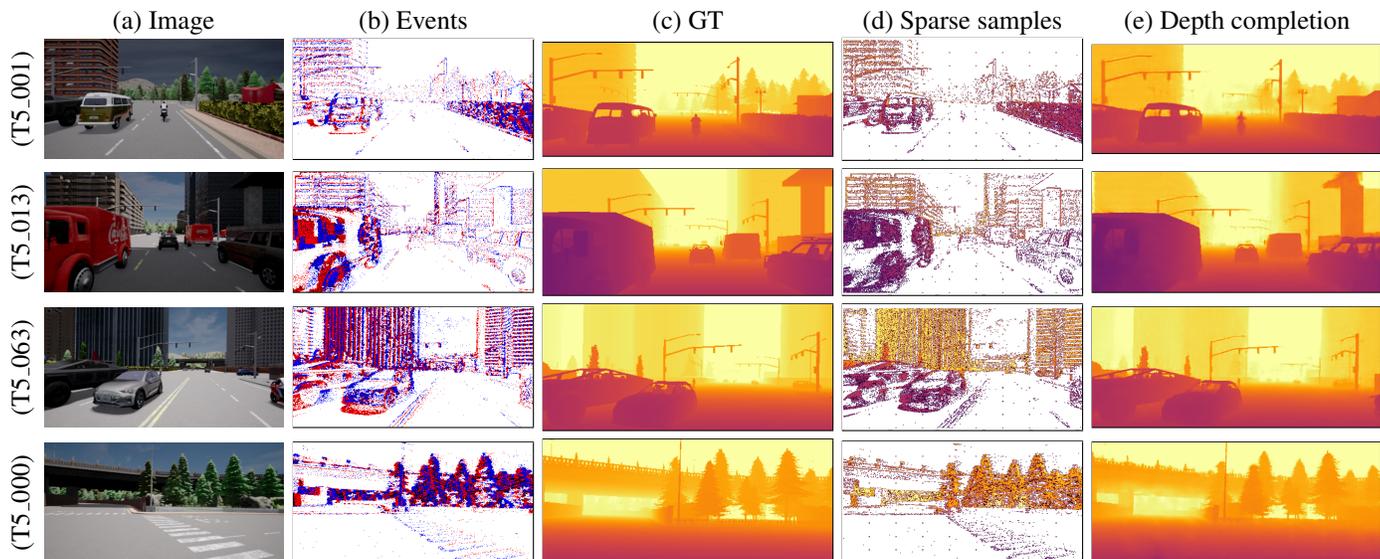


Figure 8: We evaluate our method with EventScape dataset. The dataset consists of (a) Images (b) Events and (c) GT Depth. Column (d) shows the proposed event-guided sampling and (e) our proposed depth completion network.

- future of simultaneous localization and mapping: Toward the robust-perception age,” *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, 2016.
- [11] Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder, “Foveated 3d graphics,” in *ACM SIGGRAPH Asia*, November 2012.
- [12] Diederik Paul Moeyts, Federico Corradi, Chenghan Li, Simeon A. Bamford, Luca Longinotti, Fabian F. Voigt, Stewart Berry, Gemma Taverni, Fritjof Helmchen, and Tobi Delbruck, “A sensitive dynamic and active pixel vision sensor for color or neural imaging applications,” *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, pp. 123–136, Feb. 2018.
- [13] Yuji Nozaki and Tobi Delbruck, “Temperature and parasitic photocurrent effects in dynamic vision sensors,” *IEEE Trans. Electron Devices*, vol. 64, pp. 3239–3245, Aug. 2017.
- [14] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck, “A  $128 \times 128$  120 dB 15  $\mu$ s latency asynchronous temporal contrast vision sensor,” *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.
- [15] Thomas Finatou, Atsumi Niwa, Daniel Matolin, Koya Tschimoto, Andrea Mascheroni, Etienne Reynaud, Poo-ria Mostafalu, Frederick Brady, Ludovic Chotard, Florian LeGoff, Hirotsugu Takahashi, Hayato Wakabayashi, Yusuke Oike, and Christoph Posch, “A  $1280 \times 720$  back-illuminated stacked temporal contrast event-based vision sensor with  $4.86 \mu$ m pixels, 1.066geps readout, programmable event-rate controller and compressive data-formatting pipeline,” in *IEEE Intl. Solid-State Circuits Conf. (ISSCC)*, 2020.
- [16] Hyunsurk Eric Ryu, “Industrial DVS design; key features and applications.” <http://rpg.ifi.uzh.ch/docs/CVPR19workshop/CVPRW19.Eric.Ryu.Samsung.pdf>.
- [17] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza, “Event-based motion segmentation by motion compensation,” in *Int. Conf. Comput. Vis. (ICCV)*, pp. 7243–7252, 2019.
- [18] Davide Falanga, Suseong Kim, and Davide Scaramuzza, “How fast is too fast? the role of perception latency in high-speed sense and avoid,” pp. 1884–1891, 2019.
- [19] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza, “Event-based vision: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.
- [20] Fangchang Ma, Luca Carlone, Ulas Ayaz, and Sertac Karaman, “Sparse sensing for resource-constrained depth reconstruction,” in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 96–103, 2016.
- [21] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger, “Sparsity invariant cnns,” in *2017 International Conference on 3D Vision (3DV)*, pp. 11–20, 2017.
- [22] Abdelrahman Eldesokey, Michael Felsberg, Karl Holmquist, and Michael Persson, “Uncertainty-aware cnns for depth completion: Uncertainty from beginning to end,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [23] Xinjing Cheng, Peng Wang, and Ruigang Yang, “Depth estimation via affinity learned with convolutional spatial propagation network,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [24] Fangchang Ma, Luca Carlone, Ulas Ayaz, and Sertac Karaman, “Sparse depth sensing for resource-constrained robots,” 2017.
- [25] Daniel Gehrig Javier Hidalgo-Carrio and Davide Scaramuzza, “Learning monocular dense depth from events,” *IEEE International Conference on 3D Vision (3DV)*, 2020.
- [26] Daniel Gehrig Michelle Rügge Mathias Gehrig Javier Hidalgo-Carrio and Davide Scaramuzza, “Combining events and frames using recurrent asynchronous multimodal net-

- works for monocular depth prediction,” *IEEE Robotic and Automation Letters. (RA-L)*, 2021.
- [27] Zhengqi Li and Noah Snavely, “Megadepth: Learning single-view depth prediction from internet photos,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018.
  - [28] David Eigen and Rob Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.
  - [29] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis, “Unsupervised event-based learning of optical flow, depth, and egomotion,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019.
  - [30] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza, “High speed and high dynamic range video with an event camera,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.
  - [31] Daniel Gehrig, Antonio Loquercio, Konstantinos G. Derpanis, and Davide Scaramuzza, “End-to-end learning of representations for asynchronous event-based data,” in *Int. Conf. Comput. Vis. (ICCV)*, 2019.
  - [32] Manasi Muglikar, Mathias Gehrig, Daniel Gehrig, and Davide Scaramuzza, “How to calibrate your event camera,” in *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, June 2021.