This paper has been accepted for publication at the
International Conference on 3D Vision (3DV), Online, 2021. ©IEEE

ESL: Event-based Structured Light

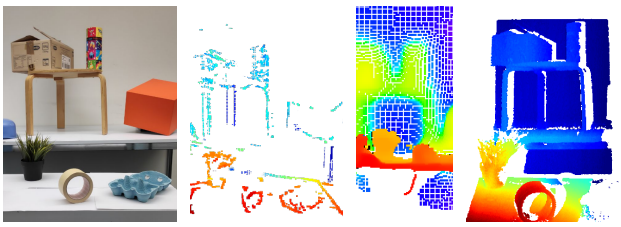
Manasi Muglikar [1] Guillermo Gallego [2] Davide Scaramuzza [1]

Abstract

Event cameras are bio-inspired sensors providing significant advantages over standard cameras such as low latency, high temporal resolution, and high dynamic range. We propose a novel structured-light system using an event camera to tackle the problem of accurate and high-speed depth sensing. Our setup consists of an event camera and a laser-point projector that uniformly illuminates the scene in a raster scanning pattern during 16 ms. Previous methods match events independently of each other, and so they deliver noisy depth estimates at high scanning speeds in the presence of signal latency and jitter. In contrast, we optimize an energy function designed to exploit event correlations, called spatio-temporal consistency. The resulting method is robust to event jitter and therefore performs better at higher scanning speeds. Experiments demonstrate that our method can deal with high-speed motion and outperform state-of-the-art 3D reconstruction methods based on event cameras, reducing the RMSE by 83% on average, for the same acquisition time. Code and dataset are available at http://rpg.ifi.uzh.ch/esl/.

1. Introduction

Depth estimation plays an important role in many computer vision and robotics applications, such as 3D modeling, augmented reality, navigation, or industrial inspection. Structured light (SL) systems estimate depth by actively projecting a known pattern on the scene and observing with a camera how light interacts (i.e., deforms and reflects) with the surfaces of the objects. In close range, these systems provide more accurate depth estimates than passive stereo methods, and so they have been used in commercial products like KinectV1 [2] and Intel RealSense [3]. Due to simple hardware and accurate depth estimates, SL systems are suitable for applications like 3D modeling, augmented reality, and indoor-autonomous navigation.



(a) 3D Scene (b) EMVS [1] (c) Intel R. (d) Ours

Figure 1: *Depth estimation with a laser point projector and an event camera.* The proposed event-based method (d) produces more dense and accurate depth estimates than (b) EMVS (events alone, without projector) or (c) a frame-based depth sensor (Intel RealSense D435). The color image (a) is not used; shown only for visualization purposes.

SL systems are constrained by the bandwidth of the devices (projector and camera) and the power of the projector light source. These constraints limit the acquisition speed, depth resolution, and performance in ambient illumination of the SL system. The main drawback of SL systems based on traditional cameras is that frame rate and redundant data acquisition limit processing to tens of Hz. By suppressing redundant data acquisition, processing can be accelerated and made more lightweight. This is a fundamental idea of recent systems based on event cameras, such as [4–6].

Event cameras, such as the Dynamic Vision Sensor (DVS) [7–9] or the ATIS sensor [10, 11], are bio-inspired sensors that measure per-pixel intensity *changes* (i.e., temporal contrast) asynchronously, at the time they occur. Thus, event cameras excel at suppressing temporal redundancy of the acquired visual signal, and they do so at the circuit level, thus consuming very little power. Moreover, event cameras have a very high temporal resolution (in the order of microseconds), which is orders of magnitude higher than that of traditional (frame-based) cameras, so they allow us to acquire visual signals at very high speed. This is another key idea of SL systems based on event cameras: the high temporal resolution simplifies data association by sequentially exposing the scene, one point [4] or one line [5] at a time. However, the unconventional output

[1]Dept. Informatics, University of Zurich and Dept. Neuroinformatics, University of Zurich and ETH Zurich, Switzerland. [2]Technische Universität Berlin and Einstein Center Digital Future, Berlin, Germany.
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.

of event cameras (a stream of asynchronous per-pixel intensity changes, called "events", instead of a synchronous sequence of images) requires the design of novel computer vision methods [12–20]. Additionally, event cameras have a very high dynamic range (HDR) (>120 dB), which allows them to operate in broad illumination conditions [21–23].

This paper tackles the problem of depth estimation using a SL system comprising a laser point-projector and an event camera (Figs. 1 and 2). Our goal is to exploit the advantages of event cameras in terms of data redundancy suppression, large bandwidth (i.e., high temporal resolution) and HDR. Early work [4] showed the potential of these types of systems; however, 3D points were estimated independently from each other, resulting in noisy 3D reconstructions. Instead, we propose to exploit the regularity of the surfaces in the world to obtain more accurate and less noisy 3D reconstructions. To this end, events are no longer processed independently, but jointly and following a forward projection model rather than the classical depth-estimation approach (stereo matching plus triangulation by back-projection).

Contributions. In summary, our contributions are:

- A novel formulation for depth estimation from an event-based SL system comprising a laser point-projector and an event camera. We model the laser point projector as an "inverse" event camera and estimate depth by maximizing the spatio-temporal consistency between the projector's and the event camera's data, when interpreted as a stereo system.
- The proposed method is robust to noise in the event timestamps (e.g., jitter, latency, BurstAER) as compared to the state-of-the-art [4].
- A convincing evaluation of the accuracy of our method using ten stationary scenes and a demonstration of the capabilities of our setup to scan eight sequences with high-speed motion.
- A dataset comprising all static and dynamic scenes recorded with our setup, and source code. To the best of our knowledge it is the first public dataset of its kind.

The following sections review the related work (Section 2), present our approach ESL (Section 3), and evaluate the method, comparing against the state-of-the-art and against ground truth data (Section 4).

2. Event-based Structured Light Systems

Prior structured light (SL) systems that have addressed the problem of depth estimation with event cameras are summarized in Table 1. Since event cameras are novel sensors (commercialized since 2008), there are only a handful of papers on SL systems. These can be classified according to whether the shape of the light source (point, line, 2D pattern) and according to the number of event cameras used.

One of the earliest works combined a DVS with a pulsed laser line to reconstruct a small terrain [5]. The pulsed laser line was projected at a fixed angle with respect to the DVS while the terrain moved beneath, perpendicular to the projected line. The method used an adaptive filter to distinguish the events caused by the laser (up to $f \sim 500$ Hz) from the events caused by noise or by the terrain's motion.

Figure 2: *Physical setup used in the experiments* (Sect. 4.1). From left to right: Intel RealSense, Prophesee Gen3S1.1 event camera and Sony Mobile point projector MPCL1A. Our method is designed for the point projector and the event camera, here separated by a 11 cm baseline. The Intel RealSense D435 sensor is only used for comparison.

Method	Event camera(s)	(pixels)	Projector
Brandli [5]	DVS128	128 × 128	Laser line 500 Hz
Matsuda [4]	DVS128	128 × 128	Laser point 60 fps
Leroux [24]	ATIS	304 × 240	DLP TI LightCrafter 3000
Mangalore [25]	DAVIS346	346 × 260	DLP TI LightCrafter 4500
Martel [6]	Stereo DAVIS240	240 × 180	Laser beam
Ours	Prophesee Gen3S1.1	640 × 480	Laser point 60 fps

Table 1: Summary of event-camera–based structured-light depth estimation works.

The SL system MC3D [4] comprised a laser point projector (operating up to 60 Hz) and a DVS. The laser raster-scanned the scene, and its reflection was captured by the DVS, which 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. In [24], a Digital Light Processing (DLP) projector was used to illuminate the scene with frequency-tagged light patterns. Each pattern's unique frequency facilitated the establishment of correspondences between the patterns and the events, leading to a sparse depth map that was later interpolated.

Recently, [6] combined a laser light source with a *stereo* setup consisting of two DAVIS event cameras [26]. 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.

More recently, [25] proposed a SL system with a fringe projector and an event camera. A sinusoidal 2D pattern with

different frequencies illuminated the scene and its reflection was captured by the camera and processed (by phase unwrapping) to generated depth estimates.

The closest work to our method is MC3D [4] since both use a laser point-projector and a single event camera, which is a sufficiently general and simple scenario that allows us to exploit the high-speed advantages of event cameras and the focusing power of a point light source. In both methods we may interpret the laser and camera as a stereo pair. The principle behind MC3D is to map the spatial disparity between the projector and event camera to temporal information of the events. When events are generated, their timestamps are mapped to disparity by multiplying by the projector's scanning speed. This operation amplifies the noise inherent in the event timestamps and leads to brittle stereo correspondences. Moreover, this noise amplification depends on the projector's speed, which is product of the projector resolution and scanning frequency. Hence, MC3D's performance degrades as the scanning frequency increases. By contrast, our method maximizes the spatio-temporal consistency between the projector's and event camera's data, thus leading to lower errors (especially with higher scanning frequencies). By exploiting the regularities in neighborhoods of event data, as opposed to the point-wise operations in MC3D, our method improves robustness against noise.

3. Depth Estimation

This section introduces basic considerations of the event-camera - projector setup (Section 3.1) and then presents our optimization approach to depth estimation (Section 3.2) using spatio-temporal consistency between the signals used on the event camera and the projector. Overall, our method is summarized in Fig. 3 and Algorithm 1.

3.1. Basic Considerations

We consider the problem of depth estimation using a laser point projector and an event camera. Fig. 3 illustrates the geometry of our configuration. The projector illuminates the scene by moving a laser light source in a raster scan fashion. The changes in illumination caused by the laser are observed by the event camera, whose pixels respond asynchronously by generating events[1]. Ideally, every camera pixel receives light from a single scene point, which is illuminated by a single location of the laser as it sweeps through the projector's pixel grid (Fig. 3). Since the light source moves in a predefined manner and the event camera and the projector are synchronized, as soon as an event is triggered, one can match it to the current light source pixel

[1]An event camera generates an event $e_k = (\mathbf{x}_k, t_k, p_k)$ at time t_k when the increment of logarithmic brightness at the pixel $\mathbf{x}_k = (x_k, y_k)^\top$ reaches a 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 Δt_k is the time since the last event at the same pixel location. [12,27]

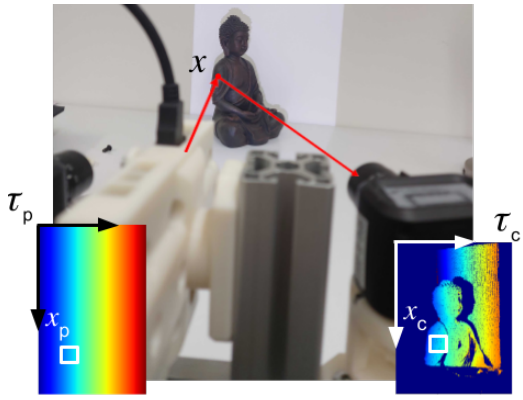


Figure 3: *Geometry of the setup*. The laser projector illuminates the scene one point $\mathbf{x}$ at a time via a ray through point $\mathbf{x}_p$, which is recorded by the camera as an event at pixel $\mathbf{x}_c$. As justified in Section 4.1, we adopt a vertical raster scan pattern (from top to bottom and left to right). The timestamps τ_p of a scan pass of the projector ($1/f$ seconds) and the timestamps of the corresponding events (τ_c) are displayed in pseudocolor, from blue (past) to red (recent). The projector does not emit color patterns.

(neglecting latency, light traveling time, etc.) to establish a stereo correspondence between the projector and the camera. This concept of converting *point-wise* temporal information into disparity was explored in [4], which relied on precise timing of both laser and event camera to establish accurate correspondences. However this one-to-one, ideal situation breaks down due to noise as the scanning speed increases. Let us introduce the time constraints and effects from both devices: projector and event camera.

Projector's Sweeping Time and Sensor's Temporal Resolution. Without loss of generality, we first assume the projector and the event camera are in a canonical stereo configuration, i.e., epipolar lines are horizontal (this can be achieved via calibration and rectification). The time that it takes for the projector to move its light source from one pixel to the next one horizontally in the raster scan, δt, is inversely proportional to the scanning frequency f and the projector's spatial resolution ($W \times H$ pixels):

$$\delta t = 1/(f\, H\, W). \tag{1}$$

For example, our projector scans at $f = 60\,\text{Hz}$ and has 1920×1080 pixels, thus it takes $1/f = 16.6\,\text{ms}$ to sweep over all its pixels, spending at most $\delta t \approx 8\,\text{ns}$ per pixel. This is considerably smaller than the temporal resolution of event cameras ($1\,\mu\text{s}$) (i.e., the camera cannot perceive the time between consecutive raster-scan pixels). To overcome this issue and be able to establish stereo correspondences along epipolar lines from time measurements, we take advantage of geometry: we *rotate* the projector by 90°, so that the raster scan is now vertical (Fig. 3). The time that it now takes for the projector to move its light source from a pixel

to its adjacent one on the still horizontal epipolar line is

$$\delta t_{\text{line}} = 1/(f\,H), \tag{2}$$

the time that it takes to sweep over one of the H raster lines. For the above projector, $\delta t_{\text{line}} \approx 15.4\,\mu\text{s}$, which is larger than the temporal resolution of most event cameras [12]. Hence, the event camera is now able to distinguish between consecutive projector pixels on the same epipolar line.

Event camera noise sources: Let us describe the different noise characteristics of event cameras that factor into this system. *Latency* is defined as the time it takes for an event to be triggered since the moment the logarithmic change of intensity exceeded the threshold. Typically this latency can range from $15\,\mu\text{s}$ to $1\,\text{ms}$. Since this affects all the timestamps equally, this can be considered as a constant offset that does not affect relative timestamps between consecutive events. *Jitter* is the random noise that appears in the timestamps. This can have a huge variance depending on the scene and the illumination conditions. *BurstAER mode*. This pixel read-out mode is very common in high resolution event cameras. It is a technique used to quickly read events from the pixel array. Instead of reading out each individual event pixel (which takes longer time for higher resolution cameras), this method reads out an entire row or group of rows together and assigns the same timestamp to all the events in these rows. This causes banding effects that appear in the event timestamps, hence they also affect the quality of the reconstructed depth map.

3.2. Maximizing Spatio-Temporal Consistency

The method in [4], (mentioned in Section 3.1, and described in Section 4.2 as baseline), computes disparity independently for each event and is, therefore, highly susceptible to noise, especially as the scanning speed increases. We now propose a method that processes events in space-time neighborhoods and exploits the regularity of surfaces present in natural scenes to improve robustness against noise and produce spatially coherent 3D reconstructions.

Time Maps: The laser projector illuminates the scene in a raster scan fashion. During one scanning interval, $T = 1/f$, the projector traverses each of its pixels $\mathbf{x}_p$ at a precise time τ_p, which allows us to define a time map over the projector's pixel grid: $\mathbf{x}_p \mapsto \tau_p(\mathbf{x}_p)$. Similarly for the event camera we can define another time map (see [28]) $\mathbf{x}_c \mapsto \tau_c(\mathbf{x}_c)$, where $\tau_c(\mathbf{x}_c)$ records the timestamp of the last event at pixel $\mathbf{x}_c$. Owing to this similarity between time maps and the fact that the projector emits light whereas the camera acquires it, we think of the projector as an "inverse" event camera. That is, the projector creates an "illumination event" $\tilde{e} = (\mathbf{x}_p, t_p, 1)$ when light at time $t = t_p$ traverses pixel $\mathbf{x}_p$. These "illumination events" are sparse, follow a raster-like pattern and are $\delta t \approx 8\,\text{ns}$ apart (1).

For simplicity, we do not make a distinction between τ_p and τ_c and refer to them as time maps (i.e., regardless of whether they are in the projector's or the event camera's image plane). Exemplary time maps are shown in Fig. 3.

Geometric Configuration: A point $\mathbf{x}_c$ on the event camera's image plane transfers onto a point $\mathbf{x}_p$ on the projector's image plane following a chain of transformations that involves the surface of the objects in the scene (Fig. 3). If we represent the surface of the objects using the depth Z with respect to the event camera, we have:

$$\mathbf{x}_p = \pi_p\big(T_{pc}\;\pi_c^{-1}\big(\mathbf{x}_c, Z(\mathbf{x}_c)\big)\big) \tag{3}$$

where π_p is the perspective projection on the projector's frame, T_{pc} is the rigid-body motion from the camera to the projector, π_c^{-1} is the inverse perspective projection of the event camera (assumed to be well-defined by a unique point of intersection between the viewing ray from the camera and the surfaces in the scene).

Time Constancy Assumption. In the above geometric configuration, the "illumination events" from the projector induce regular events on the camera. Equivalently, in terms of timestamps, the time map τ_p on the projector's image plane induces a time map τ_c on the camera's image plane:

$$\tau_c(\mathbf{x}_c) = \tau_p(\mathbf{x}_p). \tag{4}$$

This equation states a *time-consistency principle* between τ_c, τ_p, which assumes negligible travel time and photoreceptor delay [29–31], i.e., instantaneous transmission from projector to camera, as if "illumination events" and regular events were simultaneous. This time-consistency principle will play the same role that photometric consistency (e.g., the brightness constancy assumption $I_2(\mathbf{x}_2) = I_1(\mathbf{x}_1)$) plays in conventional (i.e., passive) multi-view stereo.

Disparity map from stereo matching. We formulate the problem of depth estimation using epipolar search, where we compare local neighborhoods of "illumination" and regular events (of size $W \times W \times T$) on the rectified image planes, seeking to maximize their consistency.

In terms of time maps, a neighborhood $\tau_\star(\mathbf{x}_\star, W)$, of size $W \times W$ pixels around point $\mathbf{x}_\star$, is a compact representation of the spatio-temporal neighborhood of the point $\mathbf{x}_\star$, since it not only contains spatial information but also temporal one, by definition of $\tau_\star$. Our goal becomes then to maximize consistency (4), and we do so by searching for $\tau_p(\mathbf{x}_p, W)$ (along the epipolar line) that minimizes the error

$$Z^* \doteq \arg\min_Z C(\mathbf{x}_c, Z), \tag{5}$$

$$C(\mathbf{x}_c, Z) \doteq \|\tau_c(\mathbf{x}_c, W) - \tau_p(\mathbf{x}_p, W)\|^2_{L^2(W\times W)}. \tag{6}$$

Discussion of the Approach. The temporal noise characteristics of event cameras (e.g jitter, latency, BurstAER mode, etc.) influence the quality of the obtained depth maps. The advantages of the proposed method are as follows. (*i*) *Robustness to noise* (event jitter): By considering spatio-temporal neighborhoods of events for stereo

Algorithm 1: Depth estimation by spatio-temporal consistency maximization on local neighborhoods.

Input: Time maps τ_p, τ_c during one scanning interval, and calibration (intrinsic and extrinsic parameters, T_{pc}).
Output: Depth map on the event camera image plane
Procedure:
Initialize depth map $Z(\mathbf{x}_c)$ (using epipolar search along the epipolar line in the rectified projector plane)
for *each pixel* $\mathbf{x}_c$ **do**
  Find $Z^*(\mathbf{x}_c)$ (i.e., the corresponding pixel $\mathbf{x}_p$ in (3)) that minimizes (6).
[Optional] Regularize the depth map using total variation (TV) denoising.

matching, our method becomes less susceptible to individual event's jitter than point-wise methods [4]. (*ii*) *Less data required*: Point-wise methods improve depth accuracy on static scenes by averaging depth over multiple scans [4]. Our method exploits spatial relationships between events, which makes up for temporal averaging, and therefore produces good results with less data, thus enabling better reconstructions of dynamic scenes. We may further smooth the depth maps by using a non-linear refinement step. (*iii*) *Single step stereo triangulation*: Depth parametrization and stereo matching are combined in a single step, as opposed to the classical two-step approach of first establishing correspondences and then triangulating depth like SGM or SGBM. This improves accuracy by removing triangulation errors from non-intersecting rays. (*iv*) *Trade-off controllability*: Parameter W allows us to control the quality of the estimated depth maps, with a trade-off: a small W produces fine-detailed but noisy depth maps, whereas a large W filters out noise at the expense of recovering fewer details, with (over-)smooth depth maps. Noise due to BurstAER mode or temporal resolution may affect large pixel areas. We may mitigate this type of noise by using large neighborhoods at the expense of smoothing depth discontinuities. On the downside, the method is computationally more expensive than [4], albeit it is still practical.

The pseudo-code of the method is given in Alg. 1. Overall, Alg. 1 may be interpreted as a principled non-linear method to recover depth from raw measurements, which may be initialized by a simpler method, such as [4].

4. Experiments

This section evaluates the performance of our event-based SL system for depth estimation. We first introduce the hardware setup (Section 4.1) and the baseline methods and ground truth used for comparison (Section 4.2). Then we perform experiments on static scenes to quantify the accuracy of Alg. 1, and on dynamic scenes to show its high-speed acquisition capabilities (Section 4.3).

4.1. Hardware Setup

To the best of our knowledge, there is no available dataset on which the proposed method can be tested. Therefore, we build our setup using a Prophesee event camera and a laser point source projector (Fig. 2).

Event Camera: In our setup, we use a Prophesee Gen3 camera [10, 32], with a resolution of 640×480 pixels. This sensor provides only regular events (change detection, not exposure measurement) which are used for depth estimation. We use a lens with a field of view (FOV) of $60°$.

Projector Source: We use a Sony Mobile projector MP-CL1A.The projector has a scanning speed of $60\,\mathrm{Hz}$ and a resolution of 1920×1080 pixels. During one scan (an interval of $16\,\mathrm{ms}$), the point light source moves in a raster scanning pattern. The light source consists of a Laser diode (Class 3R), of wavelength 445–$639\,\mathrm{nm}$. The event camera and the laser projector are synchronized via an external jack cable. The projector's FOV is $20°$. The projector and camera are $11\,\mathrm{cm}$ apart and their optical axes form a $26°$ angle.

Calibration: We calibrate the intrinsic parameters of the event camera using a standard calibration tool (Kalibr [33]) on the images produced after converting events to images using E2VID [22] when viewing a checkerboard pattern from different angles. We calibrate the extrinsic parameters of the camera-projector setup and the intrinsic parameters of the projector using a standard tool for SL systems [34].

4.2. Baselines and Ground Truth

Let us specify the depth estimation methods used for comparison and how ground truth depth is provided.

MC3D Baseline. We implemented the state-of-the-art method proposed in [4]. Moreover, we improved it by removing the need to scan the two end-planes of the scanning volume, which were used to linearly interpolate depth. The details are described in the supplementary material.

Due to the event jitter of the event camera and noisy correspondences (e.g., missing matches), the disparity map for a single scanning period of $16\,\mathrm{ms}$ is typically noisy and has many gaps ("holes"). Hence, we apply a median filter in post-processing (also used by [4]). However, this process does not remove all noise. Hence, we apply inpainting with hole filling and total variational (TV) denoising in post-processing. In the experiments, we use as baseline the MC3D method [4] with a single single scan ($16\,\mathrm{ms}$).

SGM Baseline. The main advantage of formulating the projector as an inverse event camera and its associated time map is that any stereo algorithm can be applied for disparity calculation between the projector' and event camera's time maps. We therefore test the Semi-Global Matching (SGM) method [35] on such timestamp maps.

Ground truth. We average the scans of MC3D over a period of $1\,\mathrm{s}$. With a frequency of $60\,\mathrm{Hz}$, this temporal averaging approach combines 60 depth scans into one.

Scene	David		Heart		Book-Duck		Plant		City of Lights		Cycle		Room		Desk-chair		Desk-books	
Mean depth	50 cm		50 cm		49 cm		70 cm		90 cm		90 cm		393.45 cm		171.92 cm		151.3 cm	
Metrics	FR ↑	RMSE ↓	FR ↑	RMSE ↓	FR ↑	RMSE ↓	FR ↑	RMSE ↓	FR ↑	RMSE ↓	FR ↑	RMSE ↓	FR ↑	RMSE ↓	FR ↑	RMSE ↓	FR ↑	RMSE ↓
MC3D	0.68	26.84	0.72	25.83	0.71	28.47	0.62	30.38	0.61	37.84	0.37	41.84	0.09	346.45	0.20	166.10	0.20	126.05
MC3D proc.	0.84	14.81	0.85	14.99	0.84	20.33	0.71	24.73	0.75	26.93	0.47	25.28	0.14	195.51	**0.54**	96.56	**0.40**	93.78
SGM	0.49	1.19	0.50	0.86	0.61	10.31	0.69	5.19	0.53	8.38	0.4	10.25	0.13	192.52	0.22	37.35	0.17	6.88
SGM proc.	0.81	1.08	0.84	**0.54**	0.80	7.30	0.82	5.21	0.75	6.76	0.58	16.31	0.15	192.23	0.22	36.94	0.20	7.30
Ours	0.94	0.50	0.96	0.57	0.85	1.43	0.89	1.98	0.80	1.23	0.65	1.19	0.26	161.36	0.38	33.82	0.30	4.69
Ours proc.	**0.96**	**0.46**	**0.98**	0.55	**0.88**	**1.40**	**0.91**	**1.97**	**0.87**	**1.17**	**0.66**	**1.15**	**0.28**	**161.34**	0.41	**33.79**	0.31	**5.14**

Table 2: *Static scenes*: RMSE (cm) and fill rate (FR, depth map completion) with respect to ground truth. (See Fig. 4).

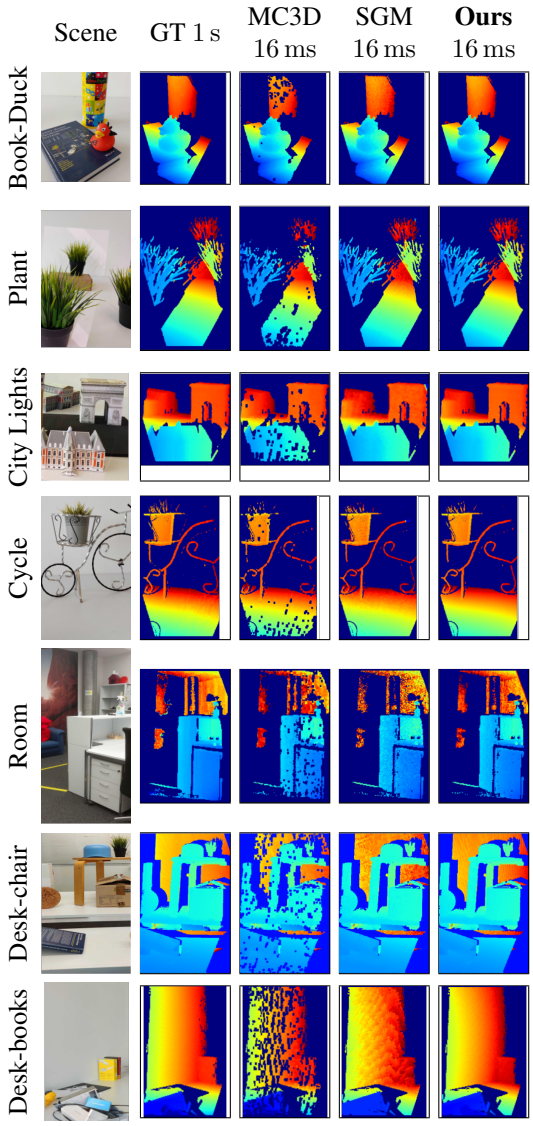


Figure 4: *Static scenes*. Depth maps produced by several SL systems on static scenes. Our method (last column) produces, with 16ms acquisition time, almost as good results as ground truth (1s temporal averaging by MC3D). The zoomed-in insets in the even rows show how the different methods deal with smooth surfaces and fine details.

Evaluation metrics. We define two evaluation metrics: (*i*) the root mean square error (*RMSE*), namely the Euclidean distance between estimates and ground truth, measured in cm, and (*ii*) the *fill rate* (or completeness), namely the percentage of ground-truth points, which have been estimated by the proposed method within a certain error. RMSE is often used to evaluate the quality of depth maps; however, this metric is heavily influenced by the scene depth, especially if there are missing points in the estimated depth map. We therefore also measure the fill rate, with a depth error threshold of 1% of the average scene depth.

4.3. Results

We assess the performance of our method on static and dynamic scenes, as well as in HDR illumination conditions.

4.3.1 Static Scenes

Static scenes enable the acquisition of accurate ground truth by temporal averaging, which ultimately allows us to assess the accuracy of our method. To this end, we evaluate our method on ten static scenes with increasing complexity: a 3D printed model of Michelangelo's David, a 3D printed model of a heart, book-duck-cylinder, plants, City of Lights and cycle-plant. We also include long-range indoor scenes of desk and room having maximum depth of 6.5 m. The scenes have varying depths (range and average depth).

Depth estimation results are collected in Fig. 4 and Table 2. The depth error was measured on the overlapping region with the ground truth. As it can be observed, on all scenes, our method, which processes the event data triggered by a single scan pass of the 60 Hz projector, outperforms the MC3D baseline method with the same input data (16 ms). Although SGM gives satisfactory results in comparison to MC3D, it suffers from artefacts that arise when temporal consistency is not strictly adhered to. Table 2 reports the fill rate (completion) and RMS error for our method and the two baselines (MC3D, SGM). The even rows incorporate post-processing ("proc"), which fills in holes (i.e., increases the fill ratio) and decreases the RMS depth error. The best results are obtained using our method and post-processing. However, the effect of post-processing is marginal in our method compared to the effect it has on the baseline methods.

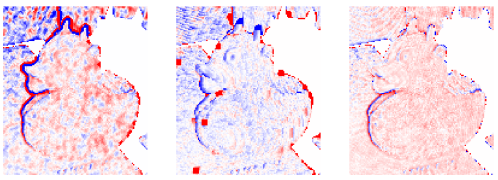


Figure 5: Signed difference depth (with respect to GT) on Book-Duck scene. Left to right: SGM, MC3D, and Ours.

Fig. 5 zooms into the signed depth errors for the Book-Duck scene (top row in Fig. 4). Here, SGM gives the largest errors, specially at the duck's edges; MC3D yields smaller errors, but still has marked object contours and gaps; finally, our approach has the smallest error contours.

4.3.2 High Dynamic Range Experiments

We also assess the performance of our method on a static scene under different illumination conditions (Fig. 6), which demonstrates the advantages of using an event-based SL depth system over conventional-camera–based depth sensor like Intel RealSense D435.

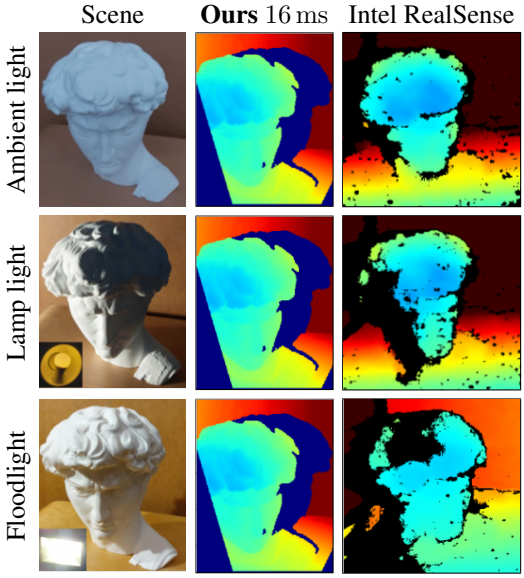


Figure 6: *Effect of varying illumination conditions* (light intensity, HDR) on depth estimation. See also Table 3.

Fig. 6 shows qualitatively how our method provides consistent depth maps under different illumination conditions, whereas a frame-based depth sensor, e.g. Intel RealSense, does not cope well with such challenging scenarios. Table 3 compares our method against the event-based baselines in HDR conditions. While all event-based methods estimate consistent depth maps across the HDR conditions, our method outperforms the MC3D baseline significantly.

Scene	Ambient light		Lamp light		Floodlight	
	FR ↑	RMSE ↓	FR ↑	RMSE ↓	FR ↑	RMSE ↓
MC3D	0.70	23.75	0.72	23.33	0.71	23.14
MC3D proc.	0.90	10.67	0.92	9.84	0.92	10.26
SGM	0.66	1.95	0.64	1.89	0.64	1.89
SGM proc.	0.90	**1.89**	0.86	**1.83**	0.86	**1.83**
Ours	**0.98**	1.99	**0.98**	1.99	**0.98**	1.99
Ours proc.	**0.98**	1.98	**0.98**	1.98	**0.98**	1.98

Table 3: *Effect of illumination conditions* on RMS error (cm) and fill rate (FR, depth map completion). (See Fig. 6).

We observe that as illumination increases, there is a slight decrease of the errors. The reason is that the noise (i.e., jitter) in the event timestamps decreases with illumination.

4.3.3 Sensitivity with respect to the Neighborhood Size

Fig. 7 qualitatively shows the performance of Alg. 1 as the size of the local aggregation neighborhood increases from $W = 3$ to $W = 15$ pixels on the event camera's image plane. As anticipated in Section 3.2, there is a trade-off between accuracy, detail preservation, and noise reduction. Our method allows us to control the desired depth estimation quality along this trade-off via the parameter W.

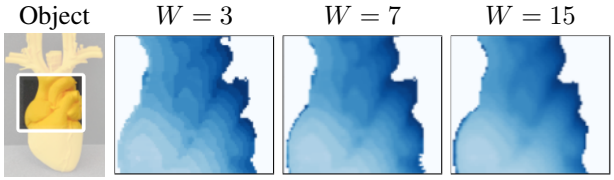


Figure 7: *Effect of the neighborhood size*: Sensitivity with respect to the size of the local aggregation neighborhood, W^2 pixels. Depth maps obtained with Alg. 1 using only one scan (16 ms), for different values of the parameter W. There is a smoothness vs. accuracy (detail preservation) trade-off that can be controlled by means of parameter W.

4.3.4 Dynamic Scenes

We also test our method on eight dynamic scenes (Fig. 8) with diverse challenging scenarios to show the capabilities of the proposed method to recover depth information in high-speed applications. Specifically, Fig. 8 shows depth recovered using our method and the baselines for the eight sequences. The figure shows a good performance of our technique in fast motion scenes and in the presence of (self-)occlusions (e.g., Scotch tape and Multi-object) and thin structures (e.g., fan). Objects do not need to be convex to recover depth with the proposed SL system. We observe that MC3D depth estimation is inaccurate due to inherent noise in the event timestamps. In the case of tape spin and fan scenes, MC3D depth has significant holes which cannot be recovered even after post-processing. SGM performs

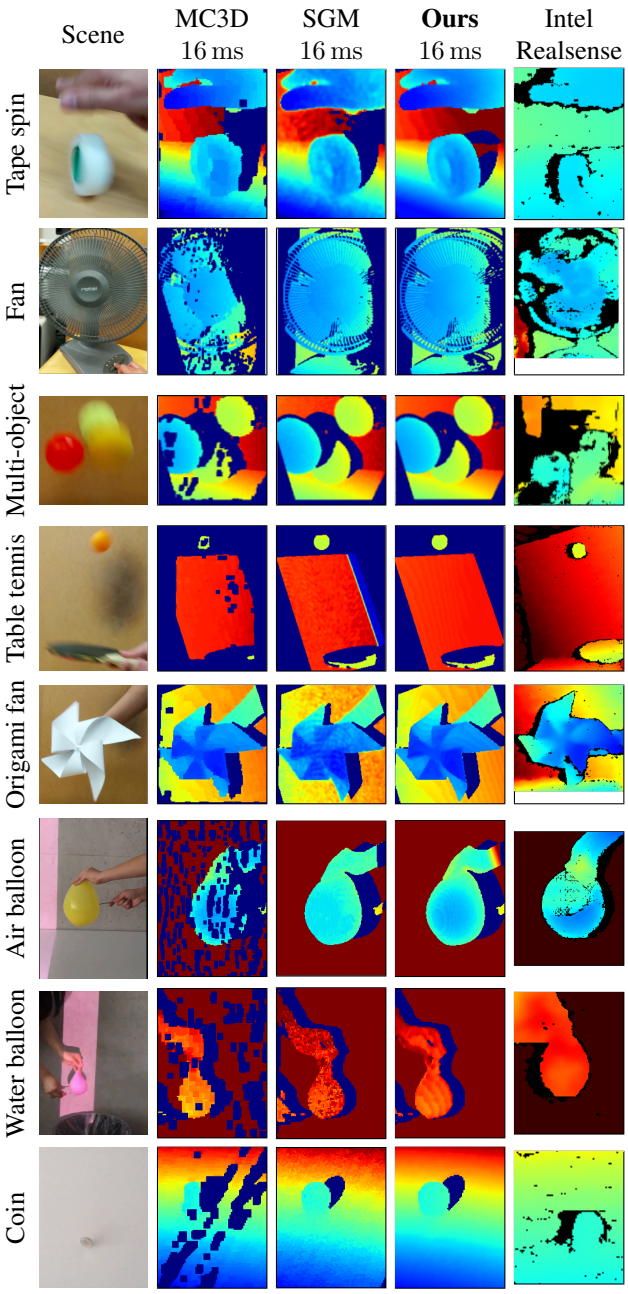


Figure 8: *Dynamic scenes*. Depth maps produced by several SL systems on dynamic scenes (spinning tape, fan, bouncing balls, popping balloons, etc.). Our method (4th column) produces the most accurate and detailed depth maps compared to the baselines.

better than MC3D, however its performance decreases in the presence of noise: in the origami fan scene, the depth along the wing of the fan and the wall has significant artefacts. Our method is robust to these artefacts and can accurately estimate depth in challenging scenes. Qualitative comparison against Intel RealSense shows favorable performance of our event-based SL method compared to frame-based SL for dynamic scenes.

Because it is *challenging* (if not impossible) to obtain *accurate* ground truth depth at ms resolution in natural dynamic scenes, such as the deforming origami fan rotating at variable speed, spinning tape, etc. we do not report quantitative results. In static scenes, we acquire accurate ground truth depth by time-averaging 1 s of scan data. However, this is not possible in dynamic scenes. The static scene experiments allow us to assess the accuracy of our method (which only requires 16 ms of data) and provide a ballpark for the accuracy of dynamic scenes.

Discussion. The experiments show that the proposed method produces, with the input data from a single scan pass, accurate depth maps at high frequency. This was possible by exploiting local event correlations at the expense of increasing the computational effort compared to MC3D. The current Python implementation of the proposed method is 38 times slower than MC3D. Nevertheless, we think this can be optimized further for real-time operation.

We also found that the method suffers in the presence of strong specularities (coin sequence, bike scene). Still, our method is able to handle specularities better than passive systems that process images using the brightness constancy assumption, which breaks down in these scenarios.

5. Conclusion

We have introduced a novel method for depth estimation using a laser point-projector and an event camera. The method aims at exploiting correlations between events (sparse space-time measurements), which previous methods on the same setup had not explored. We formulated the problem from first principles, aiming at maximizing spatio-temporal consistency while formulating the problem in an amenable stereo fashion. The experiments showed that the proposed method outperforms the frame-based (Intel RealSense) and event-based baselines (MC3D, SGM), producing, given input data from a single scan pass, similar 3D reconstruction results as the temporal average of 60 scans with MC3D. The method also provides best results in dynamic scenes and under broad illumination conditions. Exploiting local correlations was possible by introducing more event processing effort into the system. The effect of post-processing on the output of our method was marginal, signaling a thoughtful design. Finally, we think that the ideas presented here can spark a new set of techniques for high-speed depth acquisition and denoising with event-based structured light systems.

Acknowledgement

We thank Dr. Dario Brescianini and Kira Erb for their help with the prototype and data collection.

ESL: Event-based Structured Light —Supplementary Material—

Manasi Muglikar [1] Guillermo Gallego [2] Davide Scaramuzza [1]

MC3D Baseline

We implemented the state-of-the-art method proposed in [4]. Moreover, we improved it by removing the need to scan the two end-planes of the scanning volume, which were used to linearly interpolate depth, as we explain.

The method in [4] required to scan two planes at known distances from the setup at the two ends of the scanning volume. These planes were used for calibration and depth estimation. If d_n, d_f are the disparities corresponding to these two planes at depths Z_n, Z_f (near and far, respectively), then the depth Z at a pixel (x, y) with disparity $d(x, y)$ was linearly interpolated by [36]:

$$Z(x,y) = Z_n + Z_f \frac{d(x,y) - d_n(x,y)}{d_f(x,y) - d_n(x,y)} \tag{7}$$

This first-order method, which assumes pinhole models and a small illumination angle approximation throughout the scan volume, was justified in [4] to overcome the low spatial resolution of the DVS128 (128×128 pixels) and the jitter in the event timestamps.

In contrast to the setup in [4], we use a higher resolution ($\approx 20\times$) event camera and calibrate using events. Therefore, we can estimate depth from disparity without the need for prior scanning of the end-planes. In our version of MC3D, depth is given by the classical triangulation equation for a canonical stereo configuration (assuming the image planes of the projector and event camera are rectified using the calibration information):

$$Z(\mathbf{x}_c) = b\frac{F}{|\mathbf{x}_c - \mathbf{x}_p|}, \tag{8}$$

where b is the stereo baseline, F is the focal length, and the denominator is the disparity.

References

[1] Henri Rebecq, Guillermo Gallego, Elias Mueggler, and Davide Scaramuzza, "EMVS: Event-based multi-view stereo—3D reconstruction with an event camera in real-time," *Int. J. Comput. Vis.*, vol. 126, pp. 1394–1414, Dec. 2018.

[2] Microsoft Kinect v1. http://www.xbox.com/kinect.

[3] Intel RealSense Depth cameras. https://www.intelrealsense.com/coded-light/.

[4] Nathan Matsuda, Oliver Cossairt, and Mohit Gupta, "MC3D: Motion contrast 3D scanning," in *IEEE Int. Conf. Comput. Photography (ICCP)*, pp. 1–10, 2015.

[5] 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.

[6] 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.

[7] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck, "A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[8] Yunjae Suh, Seungnam Choi, Masamichi Ito, Jeongseok Kim, Youngho Lee, Jongseok Seo, Heejae Jung, Dong-Hee Yeo, Seol Namgung, Jongwoo Bong, Jun seok Kim, Paul K. J. Park, Joonseok Kim, Hyunsurk Ryu, and Yongin Park, "A 1280x960 Dynamic Vision Sensor with a 4.95-μm pixel pitch and motion artifact minimization," in *IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2020.

[9] Thomas Finateu, Atsumi Niwa, Daniel Matolin, Koya Tsuchimoto, Andrea Mascheroni, Etienne Reynaud, Pooria Mostafalu, Frederick Brady, Ludovic Chotard, Florian LeGoff, Hirotsugu Takahashi, Hayato Wakabayashi, Yusuke Oike, and Christoph Posch, "A 1280x720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86μm pixels, 1.066geps readout, programmable event-rate controller and compressive data-formatting pipeline," in *IEEE Intl. Solid-State Circuits Conf. (ISSCC)*, 2020.

[10] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt, "A QVGA 143 dB dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS," *IEEE J. Solid-State Circuits*, vol. 46, pp. 259–275, Jan. 2011.

[11] Christoph Posch, Teresa Serrano-Gotarredona, Bernabe Linares-Barranco, and Tobi Delbruck, "Retinomorphic event-based vision sensors: Bioinspired cameras with spiking output," *Proc. IEEE*, vol. 102, pp. 1470–1484, Oct. 2014.

[12] 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.

[13] Ryad Benosman, Charles Clercq, Xavier Lagorce, Sio-Hoi Ieng, and Chiara Bartolozzi, "Event-based visual flow," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 2, pp. 407–417, 2014.

[14] Hanme Kim, Ankur Handa, Ryad Benosman, Sio-Hoi Ieng, and Andrew J. Davison, "Simultaneous mosaicing and tracking with an event camera," in *British Mach. Vis. Conf. (BMVC)*, 2014.

[15] Alex Zihao Zhu, Nikolay Atanasov, and Kostas Daniilidis, "Event-based feature tracking with probabilistic data association," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 4465–4470, 2017.

[16] Henri Rebecq, Timo Horstschäfer, Guillermo Gallego, and Davide Scaramuzza, "EVO: A geometric approach to event-based 6-DOF parallel tracking and mapping in real-time," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 593–600, 2017.

[17] Guillermo Gallego, Jon E. A. Lund, Elias Mueggler, Henri Rebecq, Tobi Delbruck, and Davide Scaramuzza, "Event-based, 6-DOF camera tracking from photometric depth maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, pp. 2402–2412, Oct. 2018.

[18] Marc Osswald, Sio-Hoi Ieng, Ryad Benosman, and Giacomo Indiveri, "A spiking neural network model of 3D perception for event-based neuromorphic stereo vision systems," *Sci. Rep.*, vol. 7, Jan. 2017.

[19] Elias Mueggler, Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza, "Continuous-time visual-inertial odometry for event cameras," *IEEE Trans. Robot.*, vol. 34, pp. 1425–1440, Dec. 2018.

[20] Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza, "Focus is all you need: Loss functions for event-based vision," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 12272–12281, 2019.

[21] Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza, "Ultimate SLAM? combining events, images, and IMU for robust visual SLAM in HDR and high speed scenarios," *IEEE Robot. Autom. Lett.*, vol. 3, pp. 994–1001, Apr. 2018.

[22] 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.

[23] Gregory Cohen, Saeed Afshar, and André van Schaik, "Approaches for astrometry using event-based sensors," in *Proc. Advanced Maui Optical and Space Surveillance Technol. Conf. (AMOS)*, 2018.

[24] T. Leroux, S. H. Ieng, and R. Benosman, "Event-based structured light for depth reconstruction using frequency tagged light patterns," *arXiv e-prints*, Nov. 2018.

[25] Ashish Rao Mangalore, Chandra Sekhar Seelamantula, and Chetan Singh Thakur, "Neuromorphic fringe projection profilometry," *IEEE Signal Process. Lett.*, vol. 27, pp. 1510–1514, 2020.

[26] 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.

[27] Guillermo Gallego, Christian Forster, Elias Mueggler, and Davide Scaramuzza, "Event-based camera pose tracking using a generative event model." arXiv:1510.01972, 2015.

[28] Xavier Lagorce, Garrick Orchard, Francesco Gallupi, Bertram E. Shi, and Ryad Benosman, "HOTS: A hierarchy of event-based time-surfaces for pattern recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, pp. 1346–1359, July 2017.

[29] Yi Zhou, Guillermo Gallego, Henri Rebecq, Laurent Kneip, Hongdong Li, and Davide Scaramuzza, "Semi-dense 3D reconstruction with a stereo event camera," in *Eur. Conf. Comput. Vis. (ECCV)*, pp. 242–258, 2018.

[30] Sio-Hoi Ieng, Joao Carneiro, Marc Osswald, and Ryad Benosman, "Neuromorphic event-based generalized time-based stereovision," *Front. Neurosci.*, vol. 12, p. 442, 2018.

[31] Yi Zhou, Guillermo Gallego, and Shaojie Shen, "Event-based stereo visual odometry," *IEEE Trans. Robot.*, vol. 37, no. 5, pp. 1433–1450, 2021.

[32] Prophesee Evaluation Kits. https://www.prophesee.ai/event-based-evk/, 2020.

[33] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2013.

[34] D. Moreno and G. Taubin, "Simple, accurate, and robust projector-camera calibration," in *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization Transmission*, pp. 464–471, 2012.

[35] Heiko Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, pp. 328–341, Feb. 2008.

[36] Guijin Wang, Chenchen Feng, Xiaowei Hu, and Huazhong Yang, "Temporal matrices mapping based calibration method for event-driven structured light systems," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 1799–1808, 2021.