A Review of Event-Based Indoor Positioning and Navigation

Chenyang Shi[1], Ningfang Song[1], Wenzhuo Li[1], Yuzhen Li[1], Boyi Wei[1], Hanxiao Liu[1] and Jing Jin[1,*]

[1]*School of Instrumentation and Opto-electronics Engineering, Beihang University, Beijing, 100191, China*

Abstract

Event cameras are neuromorphic vision sensors that work differently from frame-based cameras. Instead of outputting global images of the scene at fixed frequency, event cameras generate pixel-wise output asynchronously under illumination changes. Event cameras have desirable features that make them suitable for indoor navigation and positioning: high dynamic range, high temporal resolution (consequently less motion blur) and low power consumption. However, as conventional algorithms are no longer valid for event cameras, they call for new methods to exploit their potential. This paper thus surveys sensors and algorithms for event-based navigation and positioning. We investigate event cameras (also known as Dynamic Vision Sensors) including their working principle, the trend of development and an overview of recently available sensors. We also summarize event-based algorithms that have maximized the superiority of event sensors in terms of ego-motion estimation, tracking and depth estimation. In the end, we discuss the advantages, challenges, hardware requirements and future of event cameras application in indoor navigation and positioning.

Keywords

Event camera, event-based vision, indoor positioning, indoor navigation

1. Introduction

Event cameras are bio-inspired vision sensors. They respond to the relative light changes of the natural world in an asynchronous and sparse way, completely subverting the imaging mode of the global exposure of standard camera. The event stream they output is fundamentally different from frames, boasting high time resolution, high dynamic range and low power consumption. Thus, in scenarios, event cameras are an alternative to traditional cameras. Recent studies have shown that event cameras have outperformed standard cameras in challenging positioning and mapping scenarios. Event stream naturally reflects the edges of scenes and remain low data rate, providing a new option for indoor positioning and navigation tasks that require high real-time performance. However, there are still many challenges and difficulties to be solved in practice. Therefore, we conduct a detailed investigation and discussion on the application of event cameras in positioning and navigation to further tap the potential of event cameras and provide researchers with some ideas to solve the current difficulties encountered in this field.

Currently, the main applications of event cameras are, objects detection and recognition [1, 2] feature extraction and tracking [3, 4], motion estimation [3, 5], pose estimation [6, 7], depth estimation [8, 9], video interpolation [10, 11], super-resolution [12, 13], 3D reconstruction and mapping [14, 15], etc. In the survey of [16], it reviewed the main application and development process of the event camera. Different

IPIN 2022 WiP Proceedings, September 5 - 7, 2022, Beijing, China

*Corresponding author.

shicy@buaa.edu.cn (C. Shi); Songnf@buaa.edu.cn (N. Song); 18376088@buaa.edu.cn (W. Li); 17376376@buaa.edu.cn (Y. Li); sy2017128@buaa.edu.cn (B. Wei); 19375279@buaa.edu.cn (H. Liu); jinjing@buaa.edu.cn (J. Jin)

from [16], our review aims at the application of event cameras in vision navigation and positioning, where the rationality of applying event cameras will be illustrated.

Outline: The rest of paper is organized as follows. Section.2 introduces the principle of the event cameras. Section.3 reviews the algorithms of event-based ego-motion estimation and discusses the superiority of it in complex conditions. Section.4 reviews event-based Visual Odometry (VO) and Visual Inertial Odometry (VIO) for pose estimation and tracking and discusses the development tendency of these methods. Section.5 discusses the method for event-based mapping, including depth estimation and 3D reconstruction. Section.6 summarizes the datasets for estimating the performance of these method. The paper ends with a discussion (Section.7) and a conclusion (Section.8).

2. Event Camera

Inspired by biological vision, event cameras have a completely different working mechanism compared with traditional cameras [17]. Event cameras, also known as Dynamic Vision Sensors (DVSs), no longer measure the "absolute" brightness at a constant rate, but asynchronously measure the brightness change per pixel [18, 19]. Each pixel works independently. Once the brightness change of a pixel exceeds the threshold, an event will be output at the pixel location without waiting for global exposure, which guarantees the low latency feature. Additionally, this working mechanism fundamentally gets rid of the constraint of frame rate, leading to faster response to the change of brightness (up to 1MHz) and higher dynamic range (up to $120dB$), making it capable of imaging in extremely fast motion in bright or dark environment. Moreover, as event cameras only transmit brightness changes, no output will be generated without relative displacement or change of light between the environment and the camera, which largely eliminates redundant data and reduces the transmission bandwidth and power consumption.

The output of a DVS is an event stream. Event is represented as a tuple *e = (x, y, p, t)* in which t represents the time when the pixel brightness changes in microsecond, with high sensitivity and resolution; the coordinates *(x, y)* are the position of the pixel where the brightness change occurs; polarity p indicates the direction of brightness change [20]. If the brightness increase exceeds the threshold, the polarity is +1 (ON Event). If the brightness decrease exceeds the threshold, the polarity is −1 (OFF Event).

Speicifically, DVSs refer to those whose photodiodes only contain circuits that trigger events, the mechanism of which is shown in Figure. 1. DAVISs, however, refer to those whose photodiodes can also carry out global exposure.

When V_{out} reaches the threshold V_H (upper boundary) or V_L (lower boundary), the comparator outputs the signal of OFF or ON Event, which is connected with the global exposure signal and the Address-Event-Representation (AER) handshake circuit through the OR gate. Then the row request signal is output through the handshake circuit. When the row request is answered, the column request signal is sent, and the column response signal is returned through the decision tree. This event is read out and the pixel coordinates are obtained through the address encoder.

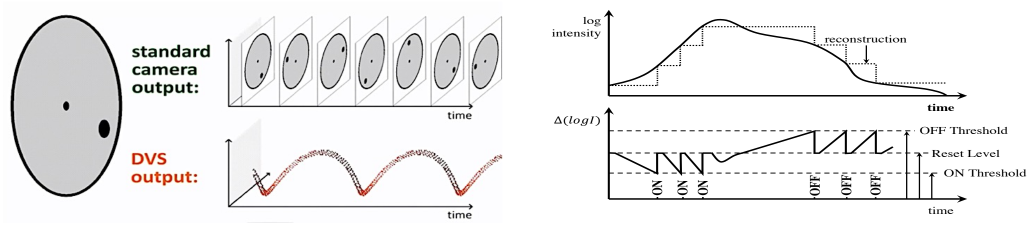
In conclusion, the superiorities of the DVS make it especially suitable for intelligent systems such as Unmanned Aerial Vehicle (UAV), aircraft, missile, intelligent shell and high-speed robot to carry out tasks such as target detection and tracking, motion estimation and autonomous navigation in indoor and outdoor environments.

With years of development, the dynamic vision sensor has made progress towards higher resolution, smaller pixel size and higher readout speed [21]. At present, the resolution of the mainstream DVS has reached 1 million pixels, with multiple modes such as gray mode, dynamic mode and optical flow mode. Table.1 introduces the parameters of the latest dynamic vision sensors in comparison to a traditional image sensor.

3. Ego-motion estmation

Ego-motion estimation recovers the state of vision sensors given their output images of the scene they are in with high accuracy. Considering the dimensions, 2D problem aims at solving 3-DOF (Degree of Freedom) motion (2-DOF translation and 1-DOF rotation or 3-DOF pure rotation) and 3D problem tackles the estimation of 6-DOF arbitrary motion.

Frame-based ego-motion estimation is mostly realized through either filter-based or optimization-based algorithms. Filter-based algorithms are the earliest applied in positioning and navigation, among



(a) The difference between imaging machanism of DVS and standard camera.

(b) The event triggering principal of DVS.

Figure 1: DVS working mechanism.

Table 1
Index of a Standard Camera and DVSs

Camera model	Blackfly S USB3	DAVIS346	DVXplorer	EB sensor [21]	DVS Gen4 [22]	CeleX-V [23]
Type	standard camera	event camera	event camera	event camera	event camera	event camera
Supplier	Teledyne FLIR	IniVation	IniVation	Prophesee	Samsung	CelePixel
Year	2020	2020	2020	2020	2019	2019
Resolution (pixel)	720×540	346×260	640×480	1280×720	1280×960	1280×800
Latency (μs)	1916	1	200	-	-	<0.5
Dynamic range (dB)	74.35	120	110	>124	90	>120
Power consumption (mW)	3000	<900	<700	73@300Meps	140	390
Pixel size (μm^2)	6.9×6.9	18.5×18.5	9×9	4.86×4.86	4.95×4.95	9.8×9.8
Max event rate	-	12	165	1066	-	100
Grayscale output?	✓	✓	×	×	×	✓

Meps: million events per second.

which the most widely used is Extended Kalman Filter (EKF). They are incremental methods, where the current camera state is considered only relevant to the camera state at one timestamp ahead. This presumption makes them suitable for small sources of data yet is rather idealized in real situations. In contrast, optimization-based algorithms are batch methods that consider all state estimation results within an interval ahead to estimate the current camera state. They concern more information and are proved to be more robust and accurate.

Event-based ego-motion estimation is carried out following two event processing patterns: (1) processing event-by-event and (2) processing on groups of events. Event-by-event-based methods enable every event to asynchronously update the system state, preserving the inherent high temporal resolution of event sensors. However, an individual event fails to depict the change of the whole scene and may suffer from strong noise signals. Therefore, it is reasonable to update the camera state with forms of event groups, such as event maps (EM), time surfaces (TSs), event frames, voxel grids and so on,

Under these patterns, the estimation problem is usually addressed within three kinds of frameworks: filter-based, optimization-based and Artificial Neural Network (ANN) -based framework. An overview of recent works on event-based ego-motion estimation can be seen in Table.2.

3.1. Filter-based framework

Probabilistic (Bayesian) filters, including Kalman filters, EKFs and particle filters (PFs), update present camera state by prior states.

Probabilistic filters have grown to be major pose estimation methods in event-by-event processing scenarios, because they naturally fit the characteristics of events: (1) filters operate asynchronous data of events, ensuring high temporal resolutions and (2) filters are particularly applicable to limited scale of computing resources of events. [24] proposed the first 6-DOF high-speed camera tracking algorithm in random natural scenes. A robust filter combining Bayesian estimation and posterior approximation of a distribution in the exponential family was put forward, enabling event-by-event pose updates from an existing photometric map of the scene. This work revealed 6-DOF high-speed tracking capabilities by event-based method and freed the tracking algorithm from limitation of scene texture.

Table 2
Event-based Method for Ego-motion Estimation

Reference	Dimension	Framework Basis	Event Representation	Scene	Open Source?	Test Dataset
Gallego [24]	3D	filter	raw event	natural	×	Gallego [24]
Chamorro [25]	3D	filter	event packets	natural	×	Chamorro [25]
Gu [26]	2D	optimization	raw event	natural	✓	Mueggler [27]
Bryner [28]	3D	optimization	event frame	natural	✓	Mueggler [27], ESIM [29], Bryner [28]
Jiao [30]	3D	optimization	TS, EM	natural	✓	ESIM [29], MVSEC [31], Zhou [32]
Peng [33]	3D	optimization	event packets	2D regularly shaped	×	UZH-FPV [34], own data
Gallego [35]	3D	optimization	IWE	natural	×	Mueggler [27]
Peng [36]	2D	optimization	IWE	natural	×	own data
Peng [37]	2D	optimization	IWE	natural	×	Mueggler [27], MVSEC [31]
Xu [38]	3D	optimization	IWE	natural	×	Mueggler [27], own data
Nunes [39]	3D	optimization	event packets	natural	✓	Mueggler [27]
Kreiser [40]	2D	ANN	raw event	natural	×	Kreiser [40]
Maqueda [41]	2D	ANN	event frame	natural	×	DDD17 [42]
Nguyen [43]	3D	ANN	event frame	natural	×	Mueggler [27]
Zhu [44]	3D	ANN	IWE	natural	×	MVSEC [31]
Kong [45]	3D	ANN	voxel grid	natural	✓	MVSEC [31], DDD17 [42], Oxford RobotCar [46]

[1] "Own data" refers to own dataset applied that is not open source.
[2] IWE: image of warped events.

In recent years, the filter-based framework has also become workable for groups of events with the contribution of event outlier rejection technique. For instance, [25] presented an EKF that updated camera pose for event packets collected within small temporal windows of $100\mu s$. This was made possible by an event-to-line matching which validated or discarded events quickly before they are stacked for estimation.

To summarize, filter-based methods suit the asynchronous nature of events and are applicable to both event-by-event and event-group algorithms. Broadly speaking, they appear to be used not as much as other methods, especially in complex scenes, due to their considerable consumption of resource to calculate and save camera states.

3.2. Optimization-based framework

Optimization is another dominant means of ego-pose estimation, which is mostly carried out on event groups. In practice, the optimization of the pose of camera is realized through the optimization of a loss function, which takes on different forms for different algorithms and optimization objectives [47].

For example, [28] tracked event camera under a maximum likelihood optimization framework from a photometric 3D map. The optimization objective is the error between the measured intensity change from event frames and the predicted intensity change calculated from the given photometric 3D map. [30] presented an enhanced motion tracker that first used TS-based method in all circumstances, and then applied EM-based method to optimize pose parameters when the optimization problem might degenerate. [33] proposed a 2D translation velocity estimation algorithm which could be seen as the backend of a VIO system. The loss function was built based on the so-called Continuous Event-Line Constraint that described the relationship between line projections from events and ego-motion of the event camera. The optimization objective was the geometric distance between the reprojected 3D line and the events.

Event-based optimization algorithms depart from conventional frame-based algorithms in that most involve motion compensation to eliminate noise and motion blur for accumulated event groups. In motion compensation algorithms, events are supposed to be triggered on the pixels which an edge moves across. [35] put forward the first unifying framework of motion compensation on the assumption that the ego-motion is uniform within a small time interval. It made a representative contribution of Contrast Maximization (CMax) framework which produced motion-compensated edge-like event images for 6-DOF camera pose estimation. It estimated the parameters of the motion that best fits a group of events by warping events to a reference time and maximizing their alignment, producing a sharp image of warped events (IWE). This fundamental framework was later renovated by several works [36, 37, 38]. [39] was another milestone that proposed the Entropy Minimization (EMin) framework. It estimated motion directly in the 3D space rather than projecting them onto image planes like CMax [35]. Therefore, it can solve motion problems in arbitrary dimensions by optimizing a family of entropy loss functions for the minimal dispersion. [26] addressed ego-motion estimation with a novel probabilistic approach which

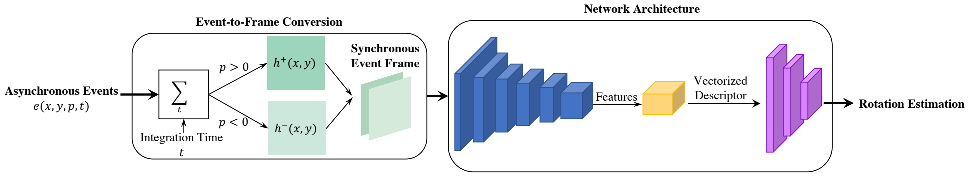


Figure 2: An example of training diagram for ANN-based pose estimation. In this example, the input of the network are synchronous event frames synthesized by integrating events in dual channels according to their polarity. Within the network, the blue layers (usually convolutional) extract features for the yellow layer to encode into a vectorized descriptor, which was sent to the purple layers for rotation estimation training.

modeled event alignment as a spatial-temporal Poisson point process. Camera rotation was estimated by maximizing joint probability of events, which achieved higher accuracy than Cmax [35], AEMin [39] and EMin [39] models in most scenarios.

In general, optimization-based methods are considered as most widely adopted for event-based ego-motion estimation. Optimization of the pose of camera is implemented by minimizing specific loss functions with the help of optimizers. Future works in this field may similar paths as prior works: refining object functions and inventing motion compensation methods for events to better depict the change of scene and upgrading existing algorithms for higher dimension of motion.

3.3. ANN-based framework

As deep-learning technologies flourish in recent years, they are popularly applied to ego-motion estimation. [48] introduced the first deep learning framework to retrieve the 6-DOF camera pose from a single frame. This groundbreaking work found that, compared to conventional key-point methods, using Convolutional Neural Network (CNN) to learn deep features appeared to be more robust in challenging scenarios such as noisy or uncleared images. This conclusion boosted the development of ANN architecture for 6-DOF pose investigation in computer vision.

Accordingly, multi-layer ANN has grown to be another mainstream method for ego-motion estimation from events, atypical structure is shown in Figure.2. They train their networks to optimize loss functions that include the state parameters of camera (also discussed in [47]). One of the representative works using event-by-event deep learning method was [40], which applied an event-based on-chip Spiking Neural Network (SNN) to the estimation of 2-DOF head pose of the iCub robot. ANN-based methods in the form of event groups are abundant to list, most of which were supervised. The work of [41] and [43] both served event frames as the neural network input, different in that [41] stacked events in dual channels of opposite polarities while [43] accumulated events in one single channel. Unlike the prior works that only used CNN or LSTM to obtain depth and geometry information, the network in [43] was composed of both a CNN to learn deep features from the event frames and a stack of LSTM to learn spatial dependencies in the image feature space, outperforming the state-of-the-art in pose estimation in general or challenging circumstances with short inference time. In recent years, unsupervised ANNs with loss functions built without restrictive conditions were also developed to solve event-based ego-motion estimation tasks. The earliest works that adopted a self and unsupervised manner still relied on input resources other than events, like greyscale images [49], or auxiliary assumptions, like the photoconsistency assumption [50]. The most recent works have realized unsupervised ANNs that only take events as input [44, 45].

To conclude, multi-layer ANNs with architectures like CNN and SNN have shown to perform well in event-based ego-motion estimation. Either original events or event groups were fed into ANNs for whom to regress the pose of camera. The birth of unsupervised networks that took pure events as input has further simplified the problem. It is prospective that novel networks will be designed to fully exploit the advantages of different architectures.

4. Event-based tracking

Estimating the pose and trajectory of rigid-body robustly and accurately is the first step to achieve positioning and mapping. Vision sensors output refined textures of the scenes for 6-DOF motion estimation. To achieve that, VO and VIO frameworks were proposed. However, under restrictions of global exposure mechanism of standard cameras, frame-based VO often suffer from motion blur especially in terms of rotation, high-speed and high-mobility motion. Adding Inertia Measurement Unit (IMU) to VO will increase the robustness of system. In a tightly or loose coupled VIO framework, triaxial angular velocity and acceleration output from IMU provide pose estimation when feature tracking fails, and visual features correct the drift of IMU. Unfortunately, when feature tracking fails for a long time, the drift cannot be corrected. In conclusion, robust vision information output is the key component of VO or VIO system.

Event cameras can output information continuously in high temporal resolution without any motion blur. Event-based VO and VIO are thus explored to deal with the problem of pose and trajectory estimation in challenging scenes.

4.1. Event-based visual odometry

Visual Odometry is a dominant approach to estimate the pose and trajectory using the visual features of scenes. If the real depth of visual features in scenes is estimated in the meanwhile, a global map can then be built, namely Simultaneously Localization and Mapping (SLAM). Similar to frame-based VO, two configuration schemes of event-based VO are generally considered, namely monocular and stereo VO. The majority of research focus on monocular event-based VO, because the configuration of this scheme is simpler than stereo scheme and comparable in terms of accuracy.

4.1.1. Angular velocity and rotation estimation

Angular velocity and rotation estimation are fundamental process of VO. An angular velocity estimation method was presented in [51], confirming that event camera was capable to estimate the 3D rotational motion of rigid-body. Currently, learning-based and optimization methods domain this field. For the learning-base methods, SNNs [52] are introduced to this task, which are comparable than ANNs-based method. For the optimization methods, CMax [53] and Rodrigues'Rotation Formula [54] are introduced as objective function for optimization.

4.1.2. Monocular and stereo visual odometry

In this task, optimization and filter-based methods are mainstreams. Filter-based methods are the first to be proposed. In [55], an event-based VO named EVO was presented, which was considered as the first SLAM system that only depended on event camera. The system transformed event streams into event frames and tracked the poses via optimizing the error between event image and semi-dense map using the inverse compositional Lucas–Kanade (L-K) method. A semi-dense 3D map was constructed by Event-based Multi-View Stereo (EMVS), a geometric 3D reconstruction method. Optimization-based methods are currently the most dominant. The choice of the objective equation for optimization is the core difference among these methods. Specifically, reprojection error minimization [56, 57], spatiotemporal registration [58] and CMax [59] et.al. It is worth noting that [59] presented an event-based VO called ETAM using continuous ray warping and volumetric contrast maximization. It extended CMax into 3D estimation, in which the target of optimization was maximizing the variance of Volume Warped Event, and achieved a sharpest warped event frame. It then built a VO, consisting of single-frame optimization as front-end based on CMax and a global optimization using B-spline curve model as back-end. In addition, there are methods [32, 60] that utilize Time Surface Maps (TSMs) to build maps and track poses while performing depth estimation.

In summary, precise pose estimation and tracking are the front-end of event-based VO, and optimization is the back-end. The key step of event-based tracking is motion compensation, the majority of aforementioned event-based VO selected optimization method to achieve that. CMax and nonlinear optimization have become mainstream in recent years, because filter-based methods occupy a large storage for saving the landmarks of a map inefficiently. Specifically, event-based tracking compute an image of warped events and sharpen the image by optimization. The sharpness of warped images reflected the accuracy of pose tracking. Thus, the objective function and tool for optimization are critical research topics. However, current event-based VO and VIO continue in the typical framework aiming at frame-based VO and VIO, the unique characteristics of event camera are still not be manifested in current processing.

4.2. Event-based visual inertial odometry

Visual Inertial Odometry is based on VO, adding IMU as a component of pose and trajectory estimation system. VIO transcends VO in terms of accuracy and robustness upon most occasions. Generally, VIO comprises of two steps, namely the front-end and the back-end. The front-end decides the data format of visual information, such as event frame and time surface. It extracts visual features as the input of the back-end. The back-end refers to the fusion method of visual information and IMU measurements, the main approaches are filter-based method [61], probabilistic method [62] and optimization method [63, 64, 65, 66]. Typically, [64] presented an approach for tightly-coupled VIO named UltimateSLAM combining events, images and IMU measurements. To synchronize the two vision sensors, this approach accumulated events as event frames at the same timestamps of standard frames and motion-compensated the event frames. It tracked the features of event frames and standard frames using FAST conner detector [67] and L-K tracker [68] individually. If the features could be triangulated and belonged to key-frames, then it fused these features and IMU measurements with nonlinear optimization, achieving the results of pose and trajectory estimation. This pipeline was demonstrated in real-time on a light-weight quadrotor system.

The original intention of adding IMU is to enhance the robustness of frame-based VO, because IMU can still maintain data output when standard cameras suffer from motion blur. The addition of the IMU has also improved the robustness of event-based VO. However, event cameras can work stably in rotating and high-speed scenes. Therefore, in theory, event cameras can perform without the addition of IMU.

5. Event-based mapping

Mapping is the final goal of SLAM. In section IV, the world coordinates of landmarks are obtained to build a sparse spatial map. However, more information of the scene is needed for diversified applications, which prompts the birth of semi-dense and dense map. Compared to sparse map, semi-dense or dense map models more or all of what is captured by the camera instead of only landmarks. These are commonly used in robot navigation, where routes and obstructions should all be reconstructed. They are also applied where 3D reconstruction with full texture of the scene or a target object is necessary for realistic and aesthetic purposes. Driven by practical purposes, this section thus focuses on semi-dense and dense mapping, which is equivalent to estimating the depth of objects in the scene. Note that mapping in most cases is preceded by ego-motion estimation, which means that the poses of cameras at all timestamps are given information.

Depth estimation in frame-based SLAM is solved in three mainstream approaches: (1) in the case of adopting a monocular camera, calculating the motion of the camera and then triangulating the depth of space points; (2) in the case of adopting a stereo camera, triangulating the depth of space points with the optical parallax between two frames; (3) using depth estimation setup, for example RGB-D camera and lidar, to directly obtain depth information. In comparison with the third approach, the former two approaches involve a significantly larger amount of computing resources and are more fragile. But they are more robust in large-scale outdoor scenes.

With event cameras emerging, event-based monocular and stereo depth estimation methods arise inheriting the former two for frame-based depth estimation. Meanwhile, event-based depth estimation using structured light is developed and works for both monocular and stereo scenarios.

5.1. Monocular depth estimation

Table. 3 lists recent event-based monocular depth estimation methods, classified according to different criteria on method and experiment. Depth estimation from a monocular event camera is a challenging task for the hardship in data association. Specifically, the temporal relationship between events cannot be straightly acquired. Therefore, early methods for event-based monocular depth estimation involved additional information, like an intensity image, in order to address the data association issue. Works in recent years have simplified the work of mapping by eliminating those auxiliary conditions.

Table 3
Event-based Method for Monocular Depth Estimation

Reference	Density	ANN-based?	Open Source?	Test Dataset
Rebecq [69]	semi-dense	×	×	Mueggler [44], own data
Gallego [35]	semi-dense	×	×	Mueggler [44]
Haessig [73]	semi-dense	✓	×	own data
Chaney [74]	semi-dense	✓	×	MVSEC [14]
Zhu [44]	semi-dense	✓	×	MVSEC [14]
Carrió [75]	dense	✓	✓	MVSEC [14]
Baudron [76]	dense	✓	✓	ShapeNet [5]
Gehrig [77]	dense	✓	✓	MVSEC [14]

"Own data" refers to own dataset applied that is not open source.

[69] did the pioneering work that reconstructed semi-dense depth map from monocular event streams without requiring event associations or intensity images. It generalized an even-based space-sweep algorithm that estimated 3D structures from frame-based MVS [70] data without traditional data association [71] for a moving event-based EMVS. In this work, individual events were considered to back-project corresponding rays that spanned spatial structures and events from multi views split the space up into disparity space image (DSI) voxels [72]. A ray counter counting rays that traversed each voxel was formed to determine ray density per voxel and a semi-dense map was obtained by computing voxels with a local maxima of ray density, which corresponds to a structural point of the scene. [35] solved the same problem as [69], estimating depth of 3D structures from multi-view events. It worked under the optimization-based CMax framework that is mentioned in Section.3.2 where events are warped into motion-corrected images. And the correct depth could be found where patches of warped events had the highest variance.

Most works in recent years addressed the mapping problem in an ANN-based fashion. However, due to the asynchronous nature of event streams, data association appeared to be a major hardship, especially for deep-learning methods. Attempts were made to achieve better events alignment by applying variable network architectures, renovating algorithms or adjusting event representation inputs. [73] transplanted the model of depth estimation from the time of focus to event-based SLAM. This work presented a novel SNN approach to the depth from defocus problem for depth map reconstruction, considering events were ideal spikes input to the SNN. The core of this network was a Leaky-Integrate-and-Fire-neurons based focus detection network composed of two input neurons for ON and OFF polarities events respectively. [74] designed an ANN specially for environments with a ground plane. It trained to learn the ratio between the height of a point from the ground plane and its depth in the event camera frame, after which height and depth information could be decomposed easily given the ground plane calibration.

Some other works discussed event representation for preserving the spatial-temporal information of event streams. CNN [44, 75, 77] is introduced for this task. [44] constructed a CNN with an unsupervised encoder-decoder architecture for depth prediction. It took discretized volumes of events as input to preserve the temporal distribution of the events as well as remove motion blur. Meanwhile, Rotational Neural Network (RNN) is introduced to handle the asynchronous data of events combined with frames. [77] did this by applying an encoder-decoder architecture on UNet which maintained an internal state that was updated asynchronously through events or frames input and could be decoded to depth estimation at any timestamp.

Overall, depth is estimated via the projection and coordinate transformation of features in monocular SLAM. Monocular depth estimation methods attempted to address the hardship of data association, which was to recover the temporal association between events. Among these methods, the ones based on deep learning have shown more robustness, because they can integrate several cues from the event stream, and thus have drawn great attention from researchers. The rise in the ability of novel methods to estimate depth from monocular events have also resulted in denser maps, which provided more detailed information of the scene for more real 3D reconstruction and more accurate navigation.

5.2. Stereo depth estimation

It is feasible to use frame-based stereo systems to estimate depth, because the shutter of the two cameras is triggered synchronously, thus feature extraction and matching for the left and right images are directly operated at the same timestamps. However, in the stereo system composed of two events cameras, the pixel matching of the left and right cameras is difficult. The principles of event co-occurrence and

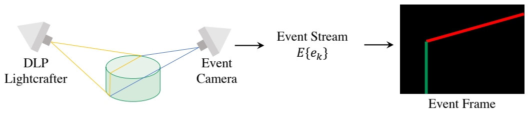


Figure 3: An example of event-driven monocular SL system. A DLP lightcrafter model casts encoded light patterns to the scene. An event camera extracts features along illuminated patterns to generate event streams. In some works, events are further aggregated into event frames to depict features in a clearer manner, where green line represents ON events and red line represents OFF events.

epipolar constraint are often used to estimate the depth. Namely, the two events triggered by the edge in 3D space are on corresponding epipolar lines of the left and right cameras. However, due to the existence of latency and noise, it is difficult to achieve this in pixel level implementation. In summary, the key step of depth estimation in event-based stereo system is finding the correspondence events of both cameras.

The most significant theory, hardship and algorithms about event-based stereo depth estimation were surveyed by [78]. For one thing, it introduced the supporting principle for stereo vision that disparity (the horizontal displacement) in two eyes of stereo camera is inversely proportional to the depth. For another, it outlined the core problem to obtaining disparity, which was to match corresponding events from two eyes along with the mismatching problem incurred by the high temporal resolution and high sensitivity of event sensors.

As was mentioned in [78], correspondence existed between disparity information and depth information in stereo problems. [79] thus realized event-based disparity estimation by introducing lifetime estimation of single events, which can be used for map reconstruction. It raised accuracy of disparity estimation by generating sharp gradient images from lifetime matching between corresponding events from two sensors. [80] utilized the velocity of event camera for generating disparity estimation. [81] developed a disparity mapping network with the stereo framework of [82] as the baseline, reserving the event embedding and stereo matching sub-networks in the previous study. In the meantime, it made major architectural modifications to the image reconstruction by integrating a cross-semantic attention mechanism and feature aggregation sub-networks by modulating event features with reconstructed image features with a stacked dilated spatially-adaptive denormalization mechanism.

In addition, window-based [83, 84], uniqueness constraint [85] and optimization method [7, 32, 86, 87, 88] were feasible for event matching. Furthermore, frame-based deep learning method [31, 86, 89, 90, 91, 92, 93, 94, 95, 96, 97] were applied to address these problems. The above works took inputs from a pair of event sensors. Distinctly, [98, 99] went down a different route. In this work, the so-called stereo setup included a frame-based camera and an event-based camera. [98] estimated dense disparity from stereo frames when they were available, predicted the disparity using odometry information, and tracked the disparity asynchronously using optical flow of events between frames.

In summary, depth is estimated via stereo matching using the disparity of two sensors in stereo SLAM. Accuracy and efficiency of events correspondence of both cameras are the key standards for evaluating stereo mapping algorithms.

5.3. Depth Estimation Using Structured Light

Structured light (SL) is considered as the most reliable technique in depth estimation. When applied in event-based SLAM, the hardware setup for an SL system mostly includes a Digital Light Process (DLP) lightcrafter model casting simple or encoded light patterns to the illuminated scene with a mirror array reflecting light back and an event camera or a pair of event cameras receiving light to generate images. A common setup for event-based monocular SL can be seen in Figure.3. Its main purpose is to simplify the extraction of features and facilitate data association in two views. In event-based systems, the measurement of spatial points using SL is accomplished by the calibration of the relative pose between the lightcrafter and the event camera, followed by triangulation when events of corresponding points are identified by data association.

A universal calibration procedure for event-driven DLP-based monocular depth estimation systems was firstly proposed by [100]. Its main contribution was a Temporal Metrices Mapping (TMM) calibration

Table 4
Resources for Event-based Positioning and Navigation

Reference	Data Volume	Motion	Scenes	Supported Tasks
Gallego [24]	10 sequences, 12GB	6-DOF	indoor/ outdoor	ego-motion estimation
Bryner [28]	12 sequences, 19GB	6-DOF	indoor/ simulated	ego-motion estimation
GRIFFIN [107]	21 sequences, 6GB	6-DOF, ornithopter	indoor/ outdoor	aerial tracking
Chamorro [25]	10 sequences, 12GB	6-DOF	indoor	tracking
EB-SLAM-3D [108]	26 sequences	6-DOF	indoor	tracking and mapping
Barranco [109]	43 sequences	translation/ rotation/ 6-DOF	indoor/ simulated	tracking
Mueggler [27]	26 sequences, 8GB	translation/ rotation/ 6-DOF, handheld	indoor/ outdoor/ simulated	tracking and mapping
UZH-FPV [34]	28 sequences, 27GB	6-DOF, drone racing	indoor/ outdoor	aggressive aerial tracking
TUM-VIE [110]	21 sequences, 447GB	translation/ 6-DOF, helmet	indoor	stereo tracking
ViVID++ [111]	36 sequences, 484GB	6-DOF, car/ handheld	indoor/ outdoor (variable illumination)	tracking and mapping
DDD17 [42]	40 sequences, 436GB	6-DOF, car	outdoor (variable illumination)	E2E ANN-based tracking
DDD20 [112]	175 sequences, 175GB	6-DOF, driving	outdoor (variable illumination)	E2E ANN-based tracking
Leung [113]	-	6-DOF, independently moving objects	indoor	ANN-based stereo tracking and sensor fusion
EVIMO [114]	>30 sequences, 30GB	6-DOF, independently moving objects	indoor	ANN-based tracking
MVSEC [31]	10 sequences, 27GB	6-DOF, car/ motorbike/ hexacopter/ handheld	indoor/ outdoor (variable illumination)	mapping
Andreopoulos [94]	12 sequences, 3GB	Rotation/ 6-DOF	indoor	mapping
DSEC [115]	53 sequences, 430GB	6-DOF, car	outdoor (variable illumination)	ANN-based mapping

E2E: end-to-end.

algorithm that calibrates the event camera and galvanometer of DLP with two temporal matrices attained through scanning a front-parallel plane and corresponding scanning speed.

As for triangulation, recent works on monocular depth estimation using SL largely focused on adopting high-frequency light patterns to fit the high temporal resolution of event cameras, such as frequency-tagged light patterns [101], blinking lights of a pseudo-random pattern [103] and periodic fringe patterns [91]. [102] projected temporally modulated lights of two wavelengths and triggered events by bispectral difference induced by light absorbance difference of a certain medium. The good merits of high temporal resolution and high dynamic range of event cameras were fully exploited to obtain unaffected bispectral difference for depth calculation. [104] built a novel formulation comprising a laser point projector and an event camera. It estimated dense depth by maximizing the spatial-temporal consistency between data from the projector and the event camera, when interpreted as a stereo system. This work took advantage of the focusing power of laser point light source and the data redundancy suppression, high temporal resolution and HDR of event camera to produce more robust mapping in high-speed motion. [105] adopted a similar hardware system with [104] but followed a more adaptive path in SL illumination where the density of projected laser in a certain area depended on the intensity of scene activity in that area to reduce power consumption.

SL can also be integrated with event-based stereo setup to simplify stereo correspondence. A typical work of event-based stereo depth estimation using SL was [106], in which a mirror-galvanometer-driven laser served as SL projector to generate blobs in the space. These blobs triggered events that were captured by two event cameras and were served as the key points for triangulation.

In general, the integration of SL has intuitively made depth features directly accessed by SLAM systems. Hardware innovations have exploited the attractive properties of events with diverse light encoding patterns adapting to the high temporal resolution merit of event cameras, while laser point light source was popularly applied to meet the HDR merit of event cameras.

6. Resources

In this section we summarize present resources (datasets and simulators) for event-based navigation and positioning, as listed in Table.4. Most of these resources have been widely applied for researchers to test the accuracy, robustness and computational efficiency of their event-based SLAM algorithms. The results served as benchmarks for the performance of new methods, which has played significant roles in driving the techniques in this field forward.

6.1. Resources for ego-motion estimation

One of the features of ego-pose estimation datasets is to provide existing information, in most cases a reconstructed depth map. [24, 28] released datasets for event-based camera tracking from an existing photometric depth map constructed by and RGB-D camera. Event streams were generated by DVS and the known photometric depth map was constructed from prior mapping by an RGB-D camera. The latter further improved the accuracy of 3D reconstructed map by attaching ElasticFusion poses from a motion capture system. [107] released the first dataset specifically for ornithopter robot perception in indoor and outdoor scenarios. This dataset was generated to prove the advantage of event camera applied to flapping-wing ornithopters.

6.2. Resources for tracking

Datasets and simulators are numerous to list for VO, VIO and SLAM. DAVISs, RGB-D cameras or stereo cameras, external motion capture systems like OptiTrack or odometry systems on hardware platforms are commonly used for generating events, depth, groundtruth motion respectively [108, 109]. One of the earliest and most classic event-based SLAM resources was [27], which released an event-based dataset and simulator for pose estimation, visual odometry and SLAM presenting variable scenes. Latter work [34, 110] boosted the study of event-based positioning and navigation with dataset derived from aggressive high-speed motions in changeable illumination scenes that were beyond the capabilities of existing tracking algorithms. In light of real-world SLAM applications, [111] proposed to include multi-sensor configurations for solving motion disturbances and illumination conditions together.

Catering to the rise of deep learning methods in event-based vision, datasets devised to train and test the performance of ANNs were released accordingly. [42] published the first annotated DAVIS driving recordings. This dataset was specially built for end-to-end (E2E) CNN and CNN/RNN networks in VO/SLAM. Vehicle speed, GPS position and driver steering, throttle, brake captured from the car's on-board diagnostics interface were given for computing ground truth. This work was expanded by [112] in terms of road types, weather and daylight conditions. Following these works, [113] published the largest event-based dataset with ground truth of independently moving entities. This dataset was recorded specially for testing deep-learning-based SLAM algorithms targeted for cameras in anomalous motion, which was made possible by including multiple labeled independently moving entities into the dataset. [114] released the first event-based dataset which included accurate pixel-wise motion masks, ego-motion and ground truth depth for the test of learning motion segmentation method.

6.3. Resources for mapping

Resources for mapping so far were mostly generated by synchronised stereo setups originally for stereo depth estimation. However, they can also be adapted to monocular depth estimation by only using events and images from one of the cameras, left or right. [31] released the first and most widely used event-based stereo depth dataset for driving, which was later improved by [115] for the first high-resolution, large-scale stereo event dataset in driving scenarios. [94] published synthetic sequences of rotating synthetic 3D object and real-world sequences of fast-rotating objects for testing the ability of algorithm to operate on nonrigid rapidly rotating objects.

7. Discussion

7.1. Current and future application of event-based positioning and navigation

In general, event camera has the ability to estimate rotation, depth, and pose in complex environments (whether indoors or outdoors) with low power consumption and without interruption. Meanwhile, it is very suitable for deployment in navigation and positioning scenarios that frequently perform complex maneuvers with strict restrictions on power consumption and highly dependency on visual information. Specifically, UAV autonomous navigation, high-speed object detection and obstacle avoidance are some examples.

Researchers will continue to work on event-based navigation and positioning algorithms that are more efficient and easier to implement on hardware. Faster and more accurate motion compensation approaches will hopefully be worked out to output high-quality poses for tracking. At the same time, in parallel pipelines, the depth of the scene can be efficiently estimated, whether monocular or stereo, and finally realize positioning and mapping in complex environments, improving robustness and speed while minimizing power consumption.

7.2. Advantages of event camera in indoor positioning and navigation

For indoor environment concerned in this paper, event camera can be incorporated as vision sensor in positioning and navigation. The high dynamic range (HDR), high temporal resolution and low power of event camera better cater to the complex characteristics of indoor environment, ensuring robust performance of the system.

7.2.1. High dynamic range

Unlike outdoor fields where natural light offers consistent illumination bright enough for camera to capture scene features, indoor environments are often dynamic with complex structures illuminated by artificial lighting. Event cameras boast high dynamic range that can reach $140dB$ compared to a common $60dB$ of frame-based cameras. This property is especially required by navigation and positioning in extreme working scenarios, for example in natural open field for long durations, where illumination condition may vary largely within a long period of time. High dynamic range ensures that navigation is consistent and robust to environmental alternations.

7.2.2. High temporal resolution

Indoor scenes are usually limited in space with a lot of obstructions. Therefore, vehicles or robots are frequently put under rapid maneuver control to avoid crashing onto obstacles. In these circumstances, the high temporal resolution property of event camera is needed for robust SLAM. Event cameras are capable of outputting event streams at microsecond level of temporal resolution in lab and sub-millisecond in the real world, enabling the navigation system to reconstruct obstacles rapidly and thus vehicles to react quickly. This also results in less motion blur as in common frame-based cameras, so that events are generated by actual features within the scene rather than noises caused by high-speed motion.

7.2.3. Low power consumption

The limited scale of indoor environment places restrictions on the volume and power dissipation of the hardware system for complicated motion of vehicle and more durable navigation. In event sensors, pixels only react to brightness changes that reach a priorly defined threshold. While the system-level power consumption of a traditional camera may be around 1 2W, that of event cameras can reach lower than 24mW. The power-saving feature makes event camera applicable to indoor onboard navigation and positioning for compact equipments that may not be able to carry power packs with mass battery.

7.3. Changllenges of event camera in indoor positioning and navigation

7.3.1. The lower bound of dynamic range

Event cameras can sense strong light intensity changes, but are not sensitive enough to weak changes (0.1lux). In extremely dim scenes, minor changes in lighting can generate large numbers of events, when in reality, they are all noise. The real events are drowned in noise, this phenomenon is very severe in low-light scenes. The latest event camera Prophesee EVK4 can perceive a minimum light level of 0.08lux and has enhanced low-light capability, but the noise problem still cannot be solved. This brings great challenges to the application of event cameras in indoor scenes that are often dimly lit. In 2021, DARPA announced that it has begun research on event cameras in the infrared band to enhance the ability of event-driven sensors to work in low-light conditions, but it still remains on paper.

7.3.2. The noise from event stream

Existing neuromorphic vision sensors suffer from three main types of output noises: background activity noise (BA), hotspot noise and flicker noise. In a static scene, most noises can be easily removed by judging the time correlation and the flicker frequency in the sliding time window. However, when the sensor performs complex motion, it is very difficult to remove hot spot noise and flicker noise. The plane is represented as a region, not a sparse point. Events generated by ambient light and reflections in windows have longer timestamps due to camera motion, much like events generated by dynamic objects in the scene. Meanwhile, events triggered by the static objects without the flickering effect under the motion of camera are temporally consistent. Using methods such as TS, it is easy to distinguish static from dynamic containing flicker noise, but it is difficult to distinguish flicker noise from real dynamic objects. By accurately estimating the camera trajectory, optical flow estimation and pixel area matching, the flicker noise from reflective objects can be judged to a certain extent. But this is still difficult in practice, because it is hard to effectively extract and track their features. We are likely to regard a mirror as a dynamic object and ignore it when building a map, which is easy to cause collisions.

7.3.3. The configuration scheme of sensors

The hardware configuration of existing event-based systems for navigation includes single event camera, binocular event cameras, an event camera combined with other visual sensor, and multi-source integration with IMU. The above schemes are proven to be feasible. From a complementary perspective, the scheme of an event camera and a standard camera combination can take into account both high-speed and low-speed scenes. On the one hand, in high-speed exercise, the camera as the main sensor provides event flow without motion blur. On the other hand, the standard camera as the main sensor provides fine scene texture characteristics. This configuration is suitable to compensate for the lack of information when event camera stays at low speed or stationary and motion blur when standard camera moving at high speed. Judging from Section.4, a single event camera can complete the feature extraction, tracking, and depth estimation. Although the addition of IMU has proven to improve the robustness of the system, we believe that a single event camera can be competent without IMU. The task of indoor positioning and navigation is actually very complicated with lot of parallel pipelines. If detection, identification and control (such as obstacle avoidance) are summarized into part of the navigation task, then the system requires the addition of standard cameras and IMU to meet the needs of diverse tasks.

7.4. Neccesity of specific event-based hardwares for indoor navigation and positioning

7.4.1. Necessity of specific event camera for positioning and navigation

In indoor tasks, sensors with low resolution are sufficient to obtain finer scene information for the limited depth of field. The reduction of sensor resolution results in the reduction of data volume and accordingly the load pressure on the back-end data processing system as well as improved computing speed. Furthermore, the noise of sensors is much higher in a dim environment (more than half of the data being noise) than in a bright scene. The reduction in sensor resolution also reduces the amount of noise. For complex indoor environments, this improvement can greatly enhance the responsiveness and maneuverability of the system.

Existing sensors output single event streams without any denoising processing, which leads to a high event rate in complex scenes. The back-end system has to use complex algorithms or hardware to perform denoising first, which leads to low efficiency. Simultaneously, due to the separation of sensors and computing hardware, data transfer process needs to be repeated. This unnecessary process also causes a lot of computational delays. Therefore, an ideal event-based sensor for positioning and navigation should have chip-level or sensor-level denoising capabilities and output high-quality data at the sensor level, which can significantly reduce the event rate and maintain the sparsity property of event stream. In order to enhance the intelligence level of the sensor, the output data should undergo a level of preprocessing to output features that can be used for tracking, namely events that have undergone feature extraction. After these preprocessing, the data output by the sensor can be directly used by the back-end, which has the significant advantages of high efficiency, sparseness and low power consumption.

7.4.2. Necessity of specific neuromorphic processing hardware for positioning and navigation

As a matter of fact, signals triggered by event-based sensor are naturally processed by an event-based processing system, namely SNNs calculating hardware. The existing calculation hardware is generally CPU, FPGA or GPU, but they are not designed for events. In order to use these setups to handle events, the events must be converted to data formats suitable for hardware, but such transformations often sacrifice the sparse and asynchronous properties of the event itself. This leads to the acceptance of only neural network methods, but without their usually heavy calculation. Those methods do not actually play the sensing advantages of event drive. Therefore, there are still some gaps compared to ordinary visual sensors in the performance indicators of actual applications. So far no SNN training mechanism that is broadly-accepted and feasible has been generated, which should hopefully facilitate the deployment and implementation of SNN networks on the hardware to truly exert the advantages of neuropsychological perception and calculation in visual navigation and positioning. In the end, the neurological sensor and the neurological calculation hardware are combined into neurological visual navigation and positioning system, which is truly high speed, high dynamic, low power consumption.

8. Conclusion

Event cameras are the representative achievement of neuromorphic vision boasting high time resolution, high dynamic range and low latency compared with standard camera. Their emergence makes applications that traditional cameras cannot handle possible, bringing a revolution to visual applications, especially in vision navigation and positioning that are full of challenges and difficulties. In this paper, we briefly introduce the principle of event cameras. Then we overviewed the research of event-based vision in navigation and positioning, including ego-motion estimation, event-based tracking, event-based mapping and datasets for estimation and analysis. Great challenges are remained in existing event-based navigation and positioning research. But challenges are opportunities. We analyze the advantages of event-based solutions, the possible improvements and research directions and make suggestions for neuromorphic hardware specialized for navigation. Finally, we put forward prospects. We hope that this paper can give researchers inspirations, so that neuromorphic vision can play a greater role in indoor navigation and positioning as well as achieve intelligent perception and calculation in complex conditions.

References

[1] C. Huang, Event-based timestamp image encoding network for human action recognition and anticipation, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–9.

[2] A. Hadviger, I. Cvišić, I. Marković, S. Vražić, I. Petrović, Feature-based event stereo visual odometry, in: 2021 European Conference on Mobile Robots (ECMR), IEEE, 2021, pp. 1–6.

[3] A. Grimaldi, V. Boutin, L. Perrinet, S.-H. Ieng, R. Benosman, A homeostatic gain control mechanism to improve event-driven object recognition, in: 2021 International Conference on Content-Based Multimedia Indexing (CBMI), IEEE, 2021, pp. 1–6.

[4] H. Cao, G. Chen, J. Xia, G. Zhuang, A. Knoll, Fusion-based feature attention gate component for vehicle detection based on event camera, IEEE Sensors Journal 21 (2021) 24540–24548.

[5] H. Akolkar, S.-H. Ieng, R. Benosman, Real-time high speed motion prediction using fast aperture-robust event-driven visual flow, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (2020) 361–372.

[6] G. Scarpellini, P. Morerio, A. Del Bue, Lifting monocular events to 3d human poses, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1358–1368.

[7] S.-H. Ieng, J. Carneiro, M. Osswald, R. Benosman, Neuromorphic event-based generalized time-based stereovision, Frontiers in Neuroscience 12 (2018) 442.

[8] Z. Chen, Q. Zheng, P. Niu, H. Tang, G. Pan, Indoor lighting estimation using an event camera, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14760–14770.

[9] N. Risi, E. Calabrese, G. Indiveri, Instantaneous stereo depth estimation of real-world stimuli with a neuromorphic stereo-vision setup, in: 2021 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2021, pp. 1–5.

[10] X. Liu, J. Guan, R. Jiang, X. Gao, S. S. Ge, Image reconstruction with event cameras based on asynchronous particle filter, in: 2022 5th International Symposium on Autonomous Systems (ISAS), IEEE, 2022, pp. 1–6.

[11] Z. Yu, Y. Zhang, D. Liu, D. Zou, X. Chen, Y. Liu, J. S. Ren, Training weakly supervised video frame interpolation with events, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14589–14598.

[12] Y. Jing, Y. Yang, X. Wang, M. Song, D. Tao, Turning frequency to resolution: Video super-resolution via event cameras, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7772–7781.

[13] L. Wang, T.-K. Kim, K.-J. Yoon, Joint framework for single image reconstruction and super-resolution with an event camera, IEEE Transactions on Pattern Analysis & Machine Intelligence (2021) 1–1.

[14] A. J. Lee, A. Kim, Eventvlad: Visual place recognition with reconstructed edges from event cameras, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2021, pp. 2247–2252.

[15] H. Cho, J. Jeong, K.-J. Yoon, Eomvs: Event-based omnidirectional multi-view stereo, IEEE Robotics and Automation Letters 6 (2021) 6709–6716.

[16] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, et al., Event-based vision: A survey, IEEE transactions on pattern analysis and machine intelligence 44 (2020) 154–180.

[17] S.-C. Liu, T. Delbruck, G. Indiveri, A. Whatley, R. Douglas, Event-based neuromorphic systems, John Wiley & Sons, 2014.

[18] T. Delbruck, R. Berner, Temporal contrast aer pixel with 0.3%-contrast event threshold, in: Proceedings of 2010 IEEE International Symposium on Circuits and Systems, IEEE, 2010, pp. 2442–2445.

[19] C. Posch, D. Matolin, R. Wohlgenannt, A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds, IEEE Journal of Solid-State Circuits 46 (2010) 259–275.

[20] C. Posch, T. Serrano-Gotarredona, B. Linares-Barranco, T. Delbruck, Retinomorphic event-based vision sensors: bioinspired cameras with spiking output, Proceedings of the IEEE 102 (2014) 1470–1484.

[21] T. Finateu, A. Niwa, D. Matolin, K. Tsuchimoto, A. Mascheroni, E. Reynaud, P. Mostafalu, F. Brady, L. Chotard, F. LeGoff, et al., 5.10 a 1280× 720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86 μm pixels, 1.066 geps readout, programmable event-rate controller and compressive data-formatting pipeline, in: 2020 IEEE International Solid-State Circuits Conference-(ISSCC), IEEE, 2020, pp. 112–114.

[22] Y. Suh, S. Choi, M. Ito, J. Kim, Y. Lee, J. Seo, H. Jung, D.-H. Yeo, S. Namgung, J. Bong, et al., A 1280× 960 dynamic vision sensor with a 4.95-μm pixel pitch and motion artifact minimization, in: 2020 IEEE international symposium on circuits and systems (ISCAS), IEEE, 2020, pp. 1–5.

[23] S. Chen, M. Guo, Live demonstration: Celex-v: A 1m pixel multi-mode event-based sensor, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2019, pp. 1682–1683.

[24] G. Gallego, J. E. Lund, E. Mueggler, H. Rebecq, T. Delbruck, D. Scaramuzza, Event-based, 6-dof camera tracking from photometric depth maps, IEEE transactions on pattern analysis and machine intelligence 40 (2017) 2402–2412.

[25] W. O. Chamorro Hernández, J. Andrade-Cetto, J. Solà Ortega, High-speed event camera tracking, in: Proceedings of the The 31st British Machine Vision Virtual Conference, 2020, pp. 1–12.

[26] C. Gu, E. Learned-Miller, D. Sheldon, G. Gallego, P. Bideau, The spatio-temporal poisson point process: A simple model for the alignment of event camera data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13495–13504.

[27] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, D. Scaramuzza, The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam, The International Journal of Robotics Research 36 (2017) 142–149.

[28] S. Bryner, G. Gallego, H. Rebecq, D. Scaramuzza, Event-based, direct camera tracking from a photometric 3d map using nonlinear optimization, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 325–331.

[29] H. Rebecq, D. Gehrig, D. Scaramuzza, Esim: an open event camera simulator, in: Conference on robot learning, PMLR, 2018, pp. 969–982.

[30] J. Jiao, H. Huang, L. Li, Z. He, Y. Zhu, M. Liu, Comparing representations in tracking for event camera-based slam, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1369–1376.

[31] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, K. Daniilidis, The multivehicle stereo event camera dataset: An event camera dataset for 3d perception, IEEE Robotics and Automation Letters 3 (2018) 2032–2039.

[32] Y. Zhou, G. Gallego, S. Shen, Event-based stereo visual odometry, IEEE Transactions on Robotics 37 (2021) 1433–1450.

[33] P. Xin, X. Wanting, Y. Jiaqi, K. Laurent, Continuous event-line constraint for closed-form velocity initialization, arXiv preprint arXiv:2109.04313 (2021).

[34] J. Delmerico, T. Cieslewski, H. Rebecq, M. Faessler, D. Scaramuzza, Are we ready for autonomous drone racing? the uzh-fpv drone racing dataset, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 6713–6719.

[35] G. Gallego, H. Rebecq, D. Scaramuzza, A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3867–3876.

[36] X. Peng, Y. Wang, L. Gao, L. Kneip, Globally-optimal event camera motion estimation, in: European Conference on Computer Vision, Springer, 2020, pp. 51–67.

[37] X. Peng, L. Gao, Y. Wang, L. Kneip, Globally-optimal contrast maximisation for event cameras, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).

[38] J. Xu, M. Jiang, L. Yu, W. Yang, W. Wang, Robust motion compensation for event cameras with smooth constraint, IEEE Transactions on Computational Imaging 6 (2020) 604–614.

[39] U. M. Nunes, Y. Demiris, Entropy minimisation framework for event-based vision model estimation, in: European Conference on Computer Vision, Springer, 2020, pp. 161–176.

[40] R. Kreiser, A. Renner, V. R. C. Leite, B. Serhan, C. Bartolozzi, A. Glover, Y. Sandamirskaya, An on-chip spiking neural network for estimation of the head pose of the icub robot, Frontiers in Neuroscience 14 (2020) 111–126.

[41] A. I. Maqueda, A. Loquercio, G. Gallego, N. García, D. Scaramuzza, Event-based vision meets deep learning on steering prediction for self-driving cars, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5419–5427.

[42] J. Binas, D. Neil, S.-C. Liu, T. Delbruck, Ddd17: End-to-end davis driving dataset, arXiv preprint arXiv:1711.01458 (2017).

[43] A. Nguyen, T.-T. Do, D. G. Caldwell, N. G. Tsagarakis, Real-time 6dof pose relocalization for event cameras with stacked spatial lstm networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[44] A. Z. Zhu, L. Yuan, K. Chaney, K. Daniilidis, Unsupervised event-based learning of optical flow, depth, and egomotion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 989–997.

[45] D. Kong, Z. Fang, K. Hou, H. Li, J. Jiang, S. Coleman, D. Kerr, Event-vpr: End-to-end weakly supervised deep network architecture for visual place recognition using event-based vision sensor, IEEE Transactions on Instrumentation and Measurement 71 (2022) 1–18.

[46] W. Maddern, G. Pascoe, C. Linegar, P. Newman, 1 year, 1000 km: The oxford robotcar dataset, The International Journal of Robotics Research 36 (2017) 3–15.

[47] G. Gallego, M. Gehrig, D. Scaramuzza, Focus is all you need: Loss functions for event-based vision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12280–12289.

[48] A. Kendall, M. Grimes, R. Cipolla, Posenet: A convolutional network for real-time 6-dof camera relocalization, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2938–2946.

[49] A. Z. Zhu, L. Yuan, K. Chaney, K. Daniilidis, Ev-flownet: Self-supervised optical flow estimation for event-based cameras, arXiv preprint arXiv:1802.06898 (2018).

[50] C. Ye, A. Mitrokhin, C. Fermüller, J. A. Yorke, Y. Aloimonos, Unsupervised learning of dense optical flow, depth and egomotion with event-based sensors, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2020, pp. 5831–5838.

[51] G. Gallego, D. Scaramuzza, Accurate angular velocity estimation with an event camera, IEEE Robotics and Automation Letters 2 (2017) 632–639.

[52] M. Gehrig, S. B. Shrestha, D. Mouritzen, D. Scaramuzza, Event-based angular velocity regression with spiking networks, in: 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2020, pp. 4195–4202.

[53] D. Liu, A. Parra, T.-J. Chin, Globally optimal contrast maximisation for event-based motion estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6349–6358.

[54] H. Kim, H. J. Kim, Real-time rotational motion estimation with contrast maximization over globally aligned events, IEEE Robotics and Automation Letters 6 (2021) 6016–6023.

[55] H. Rebecq, T. Horstschäfer, G. Gallego, D. Scaramuzza, Evo: A geometric approach to event-based 6-dof parallel tracking and mapping in real time, IEEE Robotics and Automation Letters 2 (2016) 593–600.

[56] C. Reinbacher, G. Munda, T. Pock, Real-time panoramic tracking for event cameras, in: 2017 IEEE International Conference on Computational Photography (ICCP), IEEE, 2017, pp. 1–9.

[57] D. Zhu, Z. Xu, J. Dong, C. Ye, Y. Hu, H. Su, Z. Liu, G. Chen, Neuromorphic visual odometry system for intelligent vehicle application with bio-inspired vision sensor, in: 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO), IEEE, 2019, pp. 2225–2232.

[58] D. Liu, A. Parra, T.-J. Chin, Spatiotemporal registration for event-based visual odometry, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4937–4946.

[59] Y. Wang, J. Yang, X. Peng, P. Wu, L. Gao, K. Huang, J. Chen, L. Kneip, Visual odometry with an event camera using continuous ray warping and volumetric contrast maximization, arXiv preprint arXiv:2107.03011 (2021).

[60] Y.-F. Zuo, J. Yang, J. Chen, X. Wang, Y. Wang, L. Kneip, Devo: Depth-event camera visual odometry in challenging conditions, arXiv preprint arXiv:2202.02556 (2022).

[61] A. Zihao Zhu, N. Atanasov, K. Daniilidis, Event-based visual inertial odometry, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5391–5399.

[62] E. Mueggler, G. Gallego, H. Rebecq, D. Scaramuzza, Continuous-time visual-inertial odometry for event cameras, IEEE Transactions on Robotics 34 (2018) 1425–1440.

[63] H. Rebecq, T. Horstschaefer, D. Scaramuzza, Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization (2017).

[64] A. R. Vidal, H. Rebecq, T. Horstschaefer, D. Scaramuzza, Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios, IEEE Robotics and Automation Letters 3 (2018) 994–1001.

[65] K. Nelson, Event-based visual-inertial odometry on a fixed-wing unmanned aerial vehicle (msc thesis), Air Force Institute of Technology (2019).

[66] C. Le Gentil, F. Tschopp, I. Alzugaray, T. Vidal-Calleja, R. Siegwart, J. Nieto, Idol: A framework for imu-dvs odometry using lines, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2020, pp. 5863–5870.

[67] E. Rosten, T. Drummond, Machine learning for high-speed corner detection, in: European conference on computer vision, Springer, 2006, pp. 430–443.

[68] B. D. Lucas, T. Kanade, et al., An iterative image registration technique with an application to stereo vision, volume 81, Vancouver, 1981.

[69] H. Rebecq, G. Gallego, E. Mueggler, D. Scaramuzza, Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time, International Journal of Computer Vision 126 (2018) 1394–1414.

[70] R. Szeliski, Computer vision: algorithms and applications, Springer Science & Business Media, 2010.

[71] R. T. Collins, A space-sweep approach to true multi-image matching, in: Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Ieee, 1996, pp. 358–363.

[72] R. Szeliski, P. Golland, Stereo matching with transparency and matting, in: Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), IEEE, 1998, pp. 517–524.

[73] G. Haessig, X. Berthelon, S.-H. Ieng, R. Benosman, A spiking neural network model of depth from defocus for event-based neuromorphic vision, Scientific reports 9 (2019) 1–11.

[74] K. Chaney, A. Zihao Zhu, K. Daniilidis, Learning event-based height from plane and parallax, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[75] J. Hidalgo-Carrió, D. Gehrig, D. Scaramuzza, Learning monocular dense depth from events, in: 2020 International Conference on 3D Vision (3DV), IEEE, 2020, pp. 534–542.

[76] A. Baudron, Z. W. Wang, O. Cossairt, A. K. Katsaggelos, E3d: Event-based 3d shape reconstruction, arXiv preprint arXiv:2012.05214 (2020).

[77] D. Gehrig, M. Rüegg, M. Gehrig, J. Hidalgo-Carrió, D. Scaramuzza, Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction, IEEE Robotics and Automation Letters 6 (2021) 2822–2829.

[78] L. Steffen, D. Reichard, J. Weinland, J. Kaiser, A. Roennau, R. Dillmann, Neuromorphic stereo vision: A survey of bio-inspired sensors and algorithms, Frontiers in neurorobotics 13 (2019) 28.

[79] A. Hadviger, I. Marković, I. Petrović, Stereo event lifetime and disparity estimation for dynamic vision sensors, in: 2019 European Conference on Mobile Robots (ECMR), IEEE, 2019, pp. 1–6.

[80] A. Z. Zhu, Y. Chen, K. Daniilidis, Realtime time synchronized event-based stereo, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 433–447.

[81] S. H. Ahmed, H. W. Jang, S. N. Uddin, Y. J. Jung, Deep event stereo leveraged by event-to-image translation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 882–890.

[82] S. Tulyakov, F. Fleuret, M. Kiefel, P. Gehler, M. Hirsch, Learning an event sequence embedding for dense event-based deep stereo, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1527–1537.

[83] F. Eibensteiner, H. G. Brachtendorf, J. Scharinger, Event-driven stereo vision algorithm based on silicon retina sensors, in: 2017 27th International Conference Radioelektronika (RADIOELEKTRONIKA), IEEE, 2017, pp. 1–6.

[84] E. Piatkowska, J. Kogler, N. Belbachir, M. Gelautz, Improved cooperative stereo matching for dynamic vision sensors with ground truth evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 53–60.

[85] D. Zou, F. Shi, W. Liu, J. Li, Q. Wang, P.-K. Park, C.-W. Shi, Y. J. Roh, H. E. Ryu, Robust dense depth map estimation from sparse dvs stereos, in: British Mach. Vis. Conf.(BMVC), volume 1, 2017.

[86] Z. Xie, S. Chen, G. Orchard, Event-based stereo depth estimation using belief propagation, Frontiers in neuroscience 11 (2017) 535.

[87] Y. Zhou, G. Gallego, H. Rebecq, L. Kneip, H. Li, D. Scaramuzza, Semi-dense 3d reconstruction with a stereo event camera, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 235–251.

[88] L. A. Camuñas-Mesa, T. Serrano-Gotarredona, S.-H. Ieng, R. Benosman, B. Linares-Barranco, Event-driven stereo visual tracking algorithm to solve object occlusion, IEEE transactions on neural networks and learning systems 29 (2017) 4223–4237.

[89] G. Dikov, M. Firouzi, F. Röhrbein, J. Conradt, C. Richter, Spiking cooperative stereo-matching at 2 ms latency with neuromorphic hardware, in: Conference on Biomimetic and Biohybrid Systems, Springer, 2017, pp. 119–137.

[90] M. Osswald, S.-H. Ieng, R. Benosman, G. Indiveri, A spiking neural network model of 3d perception for event-based neuromorphic stereo vision systems, Scientific reports 7 (2017) 1–12.

[91] A. R. Mangalore, C. S. Seelamantula, C. S. Thakur, Neuromorphic fringe projection profilometry, IEEE Signal Processing Letters 27 (2020) 1510–1514.

[92] L. Steffen, S. Ulbrich, A. Roennau, R. Dillmann, Multi-view 3d reconstruction with self-organizing maps on event-based data, in: 2019 19th International Conference on Advanced Robotics (ICAR), IEEE, 2019, pp. 501–508.

[93] L. Steffen, B. Hauck, J. Kaiser, J. Weinland, S. Ulbrich, D. Reichard, A. Roennau, R. Dillmann, Creating an obstacle memory through event-based stereo vision and robotic proprioception, in: 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), IEEE, 2019, pp. 1829–1836.

[94] A. Andreopoulos, H. J. Kashyap, T. K. Nayak, A. Amir, M. D. Flickner, A low power, high throughput, fully event-based stereo system, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7532–7542.

[95] G. Haessig, F. Galluppi, X. Lagorce, R. Benosman, Neuromorphic networks on the spinnaker platform, in: 2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), IEEE, 2019, pp. 86–91.

[96] J. Kaiser, J. Weinland, P. Keller, L. Steffen, J. Tieck, D. Reichard, A. Roennau, J. Conradt, R. Dillmann, Microsaccades for neuromorphic stereo vision, in: International Conference on Artificial Neural Networks, Springer, 2018, pp. 244–252.

[97] M. Domínguez-Morales, J. P. Domínguez-Morales, Á. Jiménez-Fernández, A. Linares-Barranco, G. Jiménez-Moreno, Stereo matching in address-event-representation (aer) bio-inspired binocular systems in a field-programmable gate array (fpga), Electronics 8 (2019) 410.

[98] Z. Wang, L. Pan, Y. Ng, Z. Zhuang, R. Mahony, Stereo hybrid event-frame (shef) cameras for 3d perception, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2021, pp. 9758–9764.

[99] A. Hadviger, I. Marković, I. Petrović, Stereo dense depth tracking based on optical flow using frames and events, Advanced Robotics 35 (2021) 141–152.

[100] G. Wang, C. Feng, X. Hu, H. Yang, Temporal matrices mapping-based calibration method for event-driven structured light systems, IEEE Sensors Journal 21 (2020) 1799–1808.

[101] T. Leroux, S.-H. Ieng, R. Benosman, Event-based structured light for depth reconstruction using frequency tagged light patterns, arXiv preprint arXiv:1811.10771 (2018).

[102] T. Takatani, Y. Ito, A. Ebisu, Y. Zheng, T. Aoto, Event-based bispectral photometry using temporally modulated illumination, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2021, pp. 15638–15647.

[103] X. Huang, Y. Zhang, Z. Xiong, High-speed structured light based 3d scanning using an event camera, Optics Express 29 (2021) 35864–35876.

[104] M. Muglikar, G. Gallego, D. Scaramuzza, Esl: Event-based structured light, in: 2021 International Conference on 3D Vision (3DV), IEEE, 2021, pp. 1165–1174.

[105] M. Muglikar, D. P. Moeys, D. Scaramuzza, Event guided depth sensing, in: 2021 International Conference on 3D Vision (3DV), IEEE, 2021, pp. 385–393.

[106] J. N. Martel, J. Müller, J. Conradt, Y. Sandamirskaya, An active approach to solving the stereo matching problem using event-based sensors, in: 2018 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2018, pp. 1–5.

[107] J. P. Rodríguez-Gómez, R. Tapia, J. L. Paneque, P. Grau, A. G. Eguíluz, J. R. Martínez-de Dios, A. Ollero, The griffin perception dataset: Bridging the gap between flapping-wing flight and robotic perception, IEEE Robotics and Automation Letters 6 (2021) 1066–1073.

[108] D. Weikersdorfer, D. B. Adrian, D. Cremers, J. Conradt, Event-based 3d slam with a depth-augmented dynamic vision sensor, in: 2014 IEEE international conference on robotics and automation (ICRA), IEEE, 2014, pp. 359–364.

[109] F. Barranco, C. Fermuller, Y. Aloimonos, T. Delbruck, A dataset for visual navigation with neuromorphic methods, Frontiers in neuroscience 10 (2016) 49.

[110] S. Klenk, J. Chui, N. Demmel, D. Cremers, Tum-vie: The tum stereo visual-inertial event dataset, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2021, pp. 8601–8608.

[111] A. J. Lee, Y. Cho, Y.-s. Shin, A. Kim, H. Myung, Vivid++: Vision for visibility dataset, IEEE Robotics and Automation Letters 7 (2022) 6282–6289.

[112] Y. Hu, J. Binas, D. Neil, S.-C. Liu, T. Delbruck, Ddd20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction, in: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2020, pp. 1–6.

[113] S. Leung, E. J. Shamwell, C. Maxey, W. D. Nothwang, Toward a large-scale multimodal event-based dataset for neuromorphic deep learning applications, in: Micro-and Nanotechnology Sensors, Systems, and Applications X, volume 10639, SPIE, 2018, pp. 279–288.

[114] A. Mitrokhin, C. Ye, C. Fermüller, Y. Aloimonos, T. Delbruck, Ev-imo: Motion segmentation dataset and learning pipeline for event cameras, in: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2019, pp. 6105–6112.

[115] M. Gehrig, W. Aarents, D. Gehrig, D. Scaramuzza, Dsec: A stereo event camera dataset for driving scenarios, IEEE Robotics and Automation Letters 6 (2021) 4947–4954.

[116] Z. Zhang, Microsoft kinect sensor and its effect, IEEE multimedia 19 (2012) 4–10.