

E-MOTION: A Dataset for Event-Based Scene Flow Estimation with Independent Moving Objects

Ivan Gutierrez Rodriguez^{1,2} , Julien Moreau¹ , Chiara Bartolozzi² , and Arren Glover² 

¹ Université de technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France

{igrodriguez, julien.moreau}@hds.utc.fr

² Istituto Italiano di Tecnologia, Genova, Italy

{chiara.bartolozzi, arren.glover}@iit.it

Abstract. Scene flow represents the 3D cartesian motion of points in the world and can be used for applications such as navigation in dynamic environments, human-robot interaction or non-rigid motion analysis, among others. Despite the potential benefits of event cameras for motion estimation tasks, due to their high temporal resolution and low-latency, not many works have yet addressed their use for scene flow estimation. Progress may be limited by the unconventional data adding complexity to established processing pipelines, but also due to the lack of event camera datasets with scene flow ground truth. With the aim of filling this gap, we present E-MOTION: a large and versatile dataset recorded with high-resolution event cameras suitable for depth, optical flow and scene flow estimation. E-MOTION features a number of independent moving objects for which ground truth poses and segmentation masks are also available. We release a total of 43 sequences with dense ground truth maps and poses at 200 Hz. The dataset is available at: <https://emotion.hds.utc.fr/>

Keywords: Event camera · Scene Flow · Dataset

1 Introduction

Visual perception development for robotics is still on-going, with improvements in accuracy, real-time operation, and operation in real-world conditions enabling robots to complete dangerous tasks currently performed by humans, or provide care and support for a generally aging population. Perception algorithms development is also still required to reach safety standards needed to rely on autonomous driving and reduce the impact of motor accidents. Data has become the cornerstone of modern algorithms, and real-world data is necessary to bridge the sim-to-real gap and have autonomous systems operating in the wild. Event cameras could be the technology to push visual systems towards higher temporal resolution, higher accuracy, lower latency, with overall low power usage. However, event camera datasets are rare and with ground truth available

only for specific tasks with respect to traditional vision datasets, owing to their much more recent invention and entrance to the market. Generally there are no high-resolution event camera datasets, with static scenes as well as moving objects, with multiple forms of well calibrated ground truth, and sufficient data for training and validation.

In this paper we introduce an event camera dataset with scene flow ground truth, which represents the 3D motion of visible surface points [36], typically expressed in the camera coordinate frame, and results from the composition of measurements of scene geometry, camera motion and object motion. Reporting scene flow ground truth allows the reporting of these components, which as shown in Fig. 1, provide a thorough description of a dynamic scene. Additionally, scene flow is a metric value and a physical measurement making it highly important for autonomous agents acting in the real-world tasks, such as self-driving cars and robots.

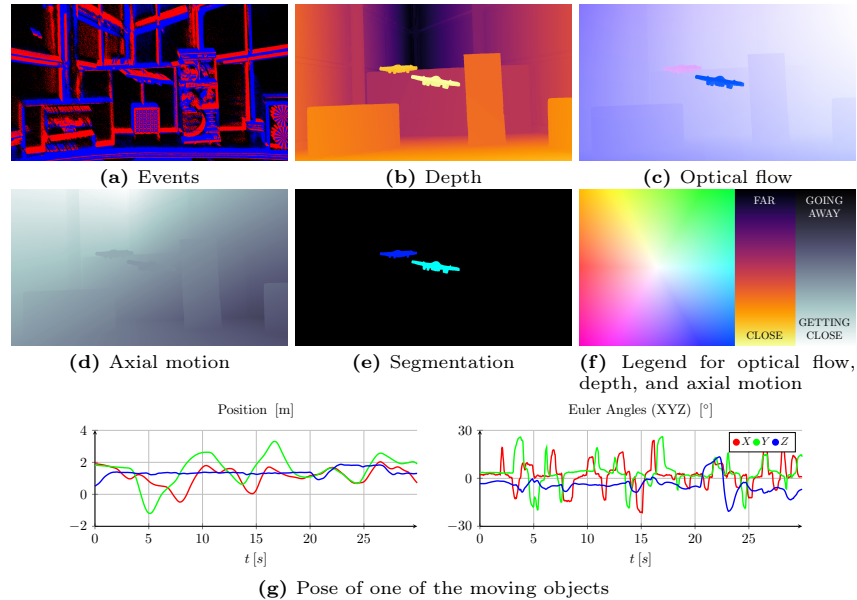


Fig. 1: Example of the data in the E-MOTION dataset. Scene flow is fully represented in 2D by depth, optical flow and axial motion maps. Axial motion is the velocity of objects in the direction of the optical axis. The legend is applicable to all figures in this paper.

Previous works explored scene flow estimation using LiDAR, RGB, and RGB-D technology [39] whereas only a few works considered the use of emerging event cameras. These relatively new visual sensors [8] offer minimum latency, very high temporal resolution, and better performance under challenging illumi-

nation conditions, which makes them perfect candidates for motion estimation tasks. Ieng et al. [19] proposed the only method for scene flow estimation with a stereo event camera. The algorithm was evaluated with simple scenes that are not publicly available. Subsequent works exploited the high temporal resolution of event streams and estimated scene flow by fusing them with LiDAR point clouds and RGB images using learning-based approaches [33,37,40]. These learning models were trained and evaluated with synthetic data due to the lack of real event camera datasets with scene flow ground truth. Datasets for depth and optical flow estimation could also be used by projecting the estimated scene flow vectors onto the image plane. However, in most of these datasets, the ground truth is generated under the assumption of a static scenario, and therefore, the presence of *Independent Moving Objects* (IMOs) is not taken into consideration (see Table 1). IMOs are important to model as they are common occurrences in real-world conditions to which scene flow algorithms should be robust. Moving objects generate important motion discontinuities which might not be properly estimated, and accounting for them leads to better trained models and more valid algorithm estimation.

We present E-MOTION: a versatile, high-resolution dataset with a stereo-pair of Prophesee EVK4 event cameras, and featuring IMOs. The data was recorded in an indoor scenario composed of different objects at different depths with some posters and calibration patterns to provide texture. The set of IMOs comprises household objects from the popular Yale-CMU-Berkeley (YCB) Object and Model set [3] and drones (quadrotors). The poses of the stereo setup and the IMOs were measured with a motion capture system synchronized with the event cameras. Together with accurate 3D models of the objects, these poses allowed us to reconstruct the scene and generate ground truth for different tasks. Overall, E-MOTION provides:

- The first dataset with real stereo event streams for scene flow evaluation.
- 2.2TB of data including 43 sequences with varied camera motions, velocities, and number of IMOs.
- Events from two calibrated high-resolution event cameras synchronized in hardware with a motion capture system.
- Spatially and temporally dense ground truth maps for depth, optical flow, scene flow and object segmentation.
- True 6DoF poses of the cameras and all the IMOs.

This paper is organized as follows: Previous event camera datasets for similar tasks are listed in section 2. The recording process, involved sensors and resulting sequences are described in section 3. Section 4 addresses the sensors synchronization and calibration. The ground truth generation process is detailed in section 5. Lastly, we discuss the usefulness of the dataset in section 6 and finish with a conclusion.

2 Previous Works

Early event camera datasets provided simple sequences for general tasks [27]. As the technology progressed, new datasets for more complex tasks were published: human pose estimation, SLAM, segmentation, etc [1]. Nevertheless, most existing datasets used early camera models, affected by noise and low spatial resolution, and only a few of them provide data from current high-resolution event cameras. In the following, we briefly survey existing datasets designed for similar tasks to E-MOTION: depth, optical flow and scene flow estimation, 6DoF object pose estimation, and segmentation of objects based on motion or appearance.

2.1 Datasets for Depth and Motion Estimation

MVSEC [42] was one of the first datasets for stereo vision with event cameras. It contains sequences recorded with DAVIS 346 cameras from various platforms. After its extension [41], ground truth depth and optical flow are available for some of the sequences. The labels were generated by fusing LiDAR data with accurate poses from a GPS or motion capture system. One of the main limitations is that scenes are assumed to be static, resulting in errors in optical flow maps where there are moving objects.

In a similar way, the autonomous driving dataset DSEC [10] provides ground truth depth computed by projecting LiDAR point clouds onto the image plane. Optical flow maps were later included by estimating camera velocity from odometry data [11]. A more recent autonomous driving dataset, CoSEC [29], provides ground truth depth and optical flow for high-resolution event cameras. In M3ED [4], depth and optical flow ground truth is also available for high-resolution event cameras, which are mounted on various platforms operating under challenging conditions. In these works, ground truth is also generated by projecting LiDAR point clouds onto the image plane. Nonetheless, all previous datasets suffer from the same problem that MVSEC: scenes are assumed to be static, resulting in ground truth inaccuracies where cars, pedestrians and other IMOs are located.

Wang et al. used VGA-resolution event cameras mounted on a robotic arm to record SHEF [38], a dataset with ground truth camera poses, derived from robot kinematics, and depth maps, generated with an accurate 3D reconstruction of the scene. In SHEF, all scenes were completely static. The EVIMO2 dataset [2] addressed the segmentation of independent moving objects also with VGA-resolution event cameras. They generated the ground truth with accurate 3D models of household objects tracked with a motion capture system. In addition to object segmentation masks, they provided depth and optical flow ground truth maps. In VECtor [9], depth ground truth was generated by projecting the scene point cloud, scanned using a FARO laser, onto the camera image plane based on the camera poses obtained with a motion capture system. The dataset uses VGA-resolution and encompasses diverse types of motion, but only static scenes.

Table 1: Comparison of our work with other datasets for depth and motion estimation with event cameras. Datasets are only considered to include IMOs when the scenario is dynamic and the motion of moving objects is taken into account in the ground truth. Ground truth frequency is not available for synthetic datasets since any value could be assumed.

Dataset	Camera			Stereo	Real	Ground truth						
	Company	Model	Resolution			IMOs	Dense	Freq. [Hz]	Dur.	Depth	Optical Flow	Scene Flow
MVSEC [42]	iniVation	DAVIS 346	346x260	✓	✓	✗	✗	20 100	43 min	✓	✓	✗
DSEC [10]	Prophesee	Gen3 CD	640x480	✓	✓	✗	✗	20	53 min	✓	✓	✗
SHEF [38]	Prophesee	Gen3 CD	640x480	✓	✓	✗	✓	125	23 min	✓	✗	✗
VECTar [9]	Prophesee	Gen3 CD	640x480	✓	✓	✗	✗	10 120	21 min	✓	✗	✗
EVIMO2 [2]	Prophesee	Gen3 CD	640x480	✓	✓	✓	✓	200	41 min	✓	✓	✗
	Samsung	DVS-Gen3	640x480	✓	✓	✓	✓	200	41 min	✓	✓	✗
M3ED [4]	Prophesee	IMX636	1280x720	✓	✓	✗	✗	10	203 min	✓	✓	✗
CoSEC [29]	Prophesee	IMX636	1280x720	✓	✓	✗	✗	≤20	60 min	✓	✓	✗
EKubric [37]	-	ESIM [30]	960x540	✗	✗	✓	✓	NA	15K fr.	✓	✓	✓
		Kubric [15]	1280x720	✗	✗	✓	✓	NA	15K fr.	✓	✓	✓
BlinkFlow [21]	-	ESIM [30]	640x480	✗	✗	✓	✓	NA	37K fr.	✗	✓	✗
		V2E [18]	640x480	✗	✗	✓	✓	NA	37K fr.	✗	✓	✗
BlinkVision [22]	-	DVS-Voltmeter [23]	640x480	✗	✗	✓	✓	NA	15K fr.	✓	✓	✓
E-MOTION (Ours)	Prophesee	IMX636	1280x720	✓	✓	✓	✓	200	42 min	✓	✓	✓

EKubric [37] was the first event camera dataset with scene flow ground truth. They use an event camera simulator [30] to generate synthetic events from high frame-rate videos and in this way adapted FlyingThings3D [24], a well known dataset for scene flow estimation with traditional cameras. Furthermore, they use Kubric [15] to generate more realistic high-resolution data. In BlinkFlow [21], more advanced event camera simulators [18, 23] were used to adapt a new high frame-rate dataset generated with Blender. The resolution was VGA and ground truth was provided for depth and optical flow. In a later extension, they released BlinkVision [22], in which scene flow ground truth is also available.

As shown in Table 1, most datasets for depth and motion estimation with event cameras disregard IMOs. Generating ground truth in the presence of IMOs is a challenging task due to the amount of information required: geometric models of all static and dynamic objects seen by the camera together with their trajectories and velocities. That is probably the reason why EKubric and BlinkVision – the only datasets with IMOs and scene flow ground truth – are synthetic. Synthetic datasets have proven to be a feasible way of generating large amounts of data to train models. Nevertheless, dealing with the sim-to-real gap requires datasets with real data.

To the best of the authors’ knowledge, E-MOTION is the first dataset recorded with real high-resolution event cameras, scene flow ground truth, and IMOs. We provide high-resolution dense ground truth maps for depth, optical flow and scene flow at 200 Hz. It is composed of 43 sequences resulting from 42 minutes of recordings with 7 different IMOs. The sequences contain different types of camera motion (translations, rotations and random), camera velocities (slow, normal and fast) and number of IMOs in the scene, allowing for the scalable development of processing methods.

2.2 Datasets for Object Pose Estimation and Tracking

Most event camera datasets providing ground truth poses for objects did not use high resolution, as shown in Table 2. Authors of RGB-DE [5] used DAVIS 346 cameras and got the ground truth by adjusting the 3D model of a toy to point clouds from an RGB-D camera. The same event camera model was used in E-POSE [17], where the ground truth was obtained from the camera poses, mounted onboard a robot with known end-effector pose; and the initial pose of the object, adjusted to a 3D reconstruction from an RGB-D camera. Another dataset using the same resolution is EVIMO [26], in which the pose of different objects was measured with a motion capture system, allowing for the simultaneous movement of both the camera and the objects. In a later extension, the authors released EVIMO2 [2], with VGA resolution and stereo event streams. High-resolution was used in YCB-Ev [31], however objects were static, placed with known position in a global coordinate frame with respect to which they could localize the camera. In EDOPT [14], resolution is close to VGA and ground truth is at 500 Hz, but however, the dataset is synthetic.

We provide ground truth poses for the 7 IMOs in E-MOTION, recorded with high-resolution event cameras. The poses are measured with a motion capture system so that there are sequences only with object motion, with both camera and object motion, and with multiple IMOs which generate occlusions with each other. Furthermore, the use of a stereo camera might allow the development of methods exploiting stereo information to improve estimations.

Table 2: Comparison of our work with other datasets for 6DoF object pose estimation and tracking with event cameras.

Dataset	Company	Camera Model	Resolution	Objects	Nº Obj.	Freq. [Hz]	Dur. [min]	Stereo	Real
RGB-DE [5]	iniVation	DAVIS 346	346x260	Dragon	1	30	2	✗	✓
YCB-Ev [31]	Prophesee	Gen4 CD	1280x720	YCB	21	90	7	✗	✓
EDOPT [14]	-	ESIM [30]	480x360	Dragon YCB	5	500	7	✗	✗
E-POSE [17]	iniVation	DAVIS 346	346x260	YCB	13	100	55	✗	✓
EVIMO [26]	iniVation	DAVIS 346	346x260	Random YCB	3	200	32	✗	✓
EVIMO2 [2]	Prophesee Samsung	Gen 3 Gen 3	640x480 640x480	Random YCB	20	200	41	✓	✓
E-MOTION (Ours)	Prophesee	IMX636	1280x720	Drones YCB	7	200	22	✓	✓

2.3 Datasets for Object Segmentation

Objects can be segmented either by their physical appearance (semantic segmentation) or by their motion when these move independently (IMO segmentation). Although object segmentation is a common computer vision task, there are not many event camera datasets providing segmentation labels for objects. DSEC was extended to provide segmentation labels of static objects [35]. The

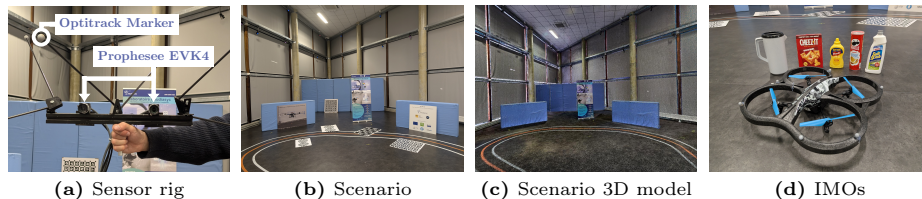


Fig. 2: Setup for recording the dataset.

Extreme Event Dataset [25] provides manually labeled data for sequences with a DAVIS240B featuring multiple moving objects. The same authors continued working on the aforementioned datasets EVIMO and EVIMO2 [26] [2], in which segmentation masks are available for all the IMOs. Nevertheless, none of these works used high-resolution cameras.

In E-MOTION we provide individual high-resolution segmentation masks for all our IMOs, which could be segmented according to both their motion and their appearance.

3 E-MOTION Dataset

The data was recorded in an indoor arena for drones, with a motion capture system and handheld event cameras. We prepared a scenario with different objects at different depths, together with some posters and calibration patterns to enrich the textures of the sequences (as shown in the picture in Fig. 2b).

3.1 Sensors

The sensors used to record the dataset are described in Table 3. Near-infrared block filters are required to reject the flashing lights of the motion capture system. The manual optics tuning and mismatch in the silicon result in slightly different bias parameters between the cameras, but they were set such that both cameras produced an approximately similar event rate. The pose of the stereo rig (Fig. 2a) and the IMOs is provided with millimeter accuracy by the Optitrack motion capture system tracking reflective markers (hereinafter referred to as optitrack). Although the sensor rig does not include an IMU, we provide a script to simulate IMU measurements from the optitrack poses with user-defined noise parameters.

3D models of all visible objects were used to reconstruct the full scene. A high-resolution point cloud was generated using a FARO laser scanner, with an accuracy on the order of millimeters and a very high-resolution to densely cover the scene, as shown in Fig. 2c. The models of the drones were obtained with a laser scanner and the photogrammetry software Meshroom [16]. YCB objects were also scanned as some official models were not up-to-date.

Table 3: Sensors used to record the dataset.

Cameras	2× Prophesee EVK4 (IMX636 sensor) 1280×720 pixels Soyo SFA0820-5M 8mm lenses ~ 41 deg FOV (overlap of 90% at 3m) Near-infrared block filters ~ 20 cm stereo baseline Biases:						
		diff	diff_off	diff_on	fo	hpf	refr
	<i>Left</i>	0	35	35	0	0	0
	<i>Right</i>	0	28	28	0	0	0
Motion Capture	Optitrack with 33 cameras eSync 2 for temporal synchronization						
Laser Scanner	FARO Focus3D						

3.2 Independent Moving Objects

As IMOs, we used a set of quadrotors and objects from the YCB dataset. We chose quadrotors – or drones – due to their ability to move freely in the space without the need of external forces. However, quadrotors were also challenging as their blades moved continuously, which introduced error in their reconstruction. The Parrot AR 2.0 Elite was specifically chosen to limit such error as the blades are partially covered.

Moving and tracking YCB objects was slightly more challenging. A thin fishing line was attached to the objects and handheld by two people on either side of the camera view, who provided motion to the objects. Additionally, the optitrack requires reflective markers for the pose measurement; to avoid changing the visual appearance of the objects, reflective tape was adhered to the back and sides of the objects with specific care in the exact placement, instead of using the standard Optitrack 3D markers. Qualitatively, we did not observe a significant change in the performance of the optitrack when using the tape instead of the 3D markers. The final set of IMOs is shown in Fig. 2d. The specific objects were chosen based on size and weight: big enough to be tracked with the optitrack while also being lightweight enough to be hung with the fishing line.

3.3 Sequences

We release 43 sequences separated in three categories: *Scene*, *Drones*, and *YCB*. In *Scene*, there are no IMOs, and the type and velocity of camera motion in each sequence is clearly defined and controlled. Camera motions can be rotations, translations, or random movements. Camera velocities can be slow, normal, or fast. In *Drones*, there is a first set of sequences in which a drone flies in front of the cameras. The camera movements are always random but the velocities are different in each sequence. In this category, we also include the case of a static camera. Moreover, we provide two sequences with two drones generating occlusions with each other. Lastly, in *YCB* we move either one or two YCB objects in front of the cameras. For each object or combination of objects, we

provide one sequence with a static camera and another with random camera motions at varying velocities.

Cameras were handheld during the recordings to allow for free and challenging movements. Albeit handheld motions are not as smooth as those that could be achieved with robots and other platforms, these may better reflect realistic and wide-ranging camera movements. Table 4 details all the sequences together with some statistics.

Table 4: Statistics of dataset sequences. TR and TE stand for train and test, respectively. All metrics are computed for the left camera, presenting minor differences with the values for the right camera.

Category	Name	Type	# IMOs	Dur. [s]	v [m/s]			ω [°/s]			Traj. length [m]	# Events [Mill.]
					AVG	STD	MAX	AVG	STD	MAX		
Scene	transl_slow_1	TR	0	100	0.11	0.04	0.27	3	2	25	10.57	782
	transl_slow_2	TR	0	97	0.16	0.07	0.40	4	2	21	14.69	829
	transl_normal_1	TR	0	77	0.26	0.13	0.62	5	3	20	19.29	939
	transl_normal_2	TR	0	53	0.41	0.21	0.95	8	4	32	19.22	924
	transl_fast_1	TR	0	31	0.90	0.46	2.09	20	11	82	19.93	962
	rot_slow_1	TR	0	110	0.02	0.01	0.06	5	3	16	1.50	936
	rot_slow_2	TR	0	107	0.02	0.01	0.07	5	2	13	1.66	909
	rot_normal_1	TR	0	40	0.06	0.04	0.18	15	8	38	2.01	832
	rot_normal_2	TR	0	47	0.06	0.03	0.15	19	8	39	2.49	1172
	rot_fast_1	TR	0	24	0.26	0.18	0.95	69	36	184	3.28	1140
	rand_slow_1	TR	0	106	0.10	0.05	0.35	8	4	30	10.64	1515
	rand_slow_2	TE	0	82	0.13	0.07	0.31	10	5	30	10.14	1360
	rand_normal_1	TE	0	77	0.30	0.16	0.78	26	14	71	20.65	2422
	rand_normal_2	TR	0	74	0.34	0.15	0.84	28	16	80	22.26	2467
	rand_fast_1	TE	0	42	0.67	0.30	1.46	64	35	199	22.07	2158
	rand_fast_2	TR	0	37	0.76	0.37	1.78	59	34	159	18.16	1768
	square_traj_1	TE	0	45	0.27	0.10	0.51	19	8	44	11.46	1121
	square_traj_2	TR	0	43	0.25	0.08	0.45	14	6	34	10.40	837
Drones	drone_static	TR	1	38	0.00	0.00	0.00	0	0	0	0.00	20
	drone_slow_1	TE	1	39	0.14	0.07	0.73	18	8	60	4.83	860
	drone_slow_2	TR	1	65	0.18	0.10	0.94	15	8	72	9.85	1549
	drone_normal_1	TE	1	45	0.32	0.17	1.47	28	15	116	12.57	1230
	drone_normal_2	TR	1	43	0.28	0.14	1.06	24	12	74	10.65	1096
	drone_fast_1	TE	1	40	0.64	0.38	1.67	73	41	202	17.43	1825
	drone_fast_2	TE	1	34	0.72	0.41	2.59	69	39	289	17.07	1513
	looping_static	TR	1	23	0.00	0.00	0.00	0	0	0	0.00	15
	looping_moving	TR	1	19	0.23	0.14	0.73	20	8	56	3.40	519
	two_drones_1	TE	2	73	0.23	0.14	0.79	22	11	67	14.34	1890
	two_drones_2	TR	2	73	0.26	0.13	1.41	25	10	57	16.48	1979
	square_traj_drone	TR	1	39	0.33	0.09	0.55	17	8	42	12.40	858
YCB	bleach_cleanser_static	TR	1	31	0.00	0.00	0.00	0	0	0	0.00	64
	bleach_cleanser	TR	1	69	0.32	0.14	0.85	29	16	99	16.58	1737
	mustard_bottle_static	TR	1	73	0.00	0.00	0.00	0	0	0	0.00	67
	mustard_bottle	TR	1	136	0.36	0.21	1.84	30	20	246	10.44	3304
	chips_can_static	TR	1	58	0.00	0.00	0.00	0	0	0	0.00	42
	chips_can	TR	1	60	0.30	0.15	0.88	26	13	96	13.20	1328
	cracker_box_static	TR	1	66	0.00	0.00	0.00	0	0	0	0.00	108
	cracker_box	TR	1	60	0.19	0.11	0.51	26	11	69	5.77	1380
	pitcher_static	TR	1	56	0.00	0.00	0.00	0	0	0	0.00	35
	mult_static_1	TR	2	32	0.00	0.00	0.00	0	0	0	0.00	63
	mult_1	TR	2	52	0.34	0.14	0.84	20	10	68	12.19	1161
	mult_static_2	TR	2	60	0.00	0.00	0.00	0	0	0	0.00	150
	mult_2	TE	2	60	0.35	0.16	1.46	19	11	105	17.28	1199

3.4 Data Format

Data is stored in HDF5³ files to maintain consistency with other major event camera datasets (e.g. [4,10]). For each sequence, we provide a file with the events from both cameras and another with the ground truth for both real and rectified

³ <https://www.hdfgroup.org/solutions/hdf5/>

cameras. Events are stored in a format inspired by DSEC [10]. Both files include calibrations, rectification matrices, and all relevant metadata.

4 Synchronization and Calibration

Our sensors are synchronized in hardware using an Optitrack eSync 2 device, which generates a square signal when data is being recorded. A pulse is generated every time the optitrack captures the pose of the objects in the scene. Each edge of the signal is received and stored by the cameras as a special event time stamped with the camera clock. Then, the three clocks (optitrack, left camera, and right camera) are synchronized by aligning the stamps of the special events with the stamps of the optitrack measurements. The data from the event cameras is recorded using a custom driver developed with the Prophesee Metavision SDK⁴.

To calibrate the cameras, we use image reconstruction [34] to allow the detection of an AprilGrid [28] pattern (P). Camera intrinsic (θ_L, θ_R) and extrinsic (${}^L T_R$) calibration parameters are estimated with Kalibr [6] using the pinhole camera model and radial-tangential distortion model (two radial and two tangential distortion parameters). In addition to the extrinsic calibration of the cameras, the rest of spatial transformations in Fig. 3a are required. The optitrack provides the pose of the cameras and the IMOs w.r.t its own coordinate frame (O). These poses correspond to rigid bodies attached to the cameras (H) and the IMOs (I_i) but not to the camera sensors (L, R) and the IMO models (M_i) themselves. Thus, the transformation from these rigid bodies to the actual camera sensors and object models must be estimated. In the same way, the optitrack coordinate frame is not intrinsically aligned with the coordinate frame of scenario model (M), which coincides with the origin of the point cloud.

Obtaining ${}^H T_L$ corresponds to the hand-eye calibration problem so that an initial estimate can be computed with the method in [7]. During hand-eye calibration, an estimate of the calibration pattern’s pose ${}^O T_P$ is obtained, and therefore, the 3D position of its corners in O . For each reconstructed image, we obtain a ground truth projection of these corners and an estimate of the camera pose, which is linked to the output of the optitrack (${}^O T_H$) and ${}^H T_L$. Combined, the process gives a set of n points $P_i({}^O T_P, {}^H T_L) \in \mathbb{R}^3$ in L with ground truth projection $p_i \in \mathbb{R}^2$. The transformations involved in the computation of the set of P_i points are represented in Fig. 3b. With these points, we can evaluate the calibration by computing the average re-projection error:

$$f({}^O T_P, {}^H T_L, \theta_L) = \sum_{i=0}^{n-1} \frac{\|p_i - \pi(P_i({}^O T_P, {}^H T_L), \theta_L)\|}{n} \quad (1)$$

where $\pi(P, \theta)$ projects a point $P \in \mathbb{R}^3$ onto the image plane of a camera with intrinsic parameters θ .

⁴ <https://docs.prophesee.ai/stable/index.html>

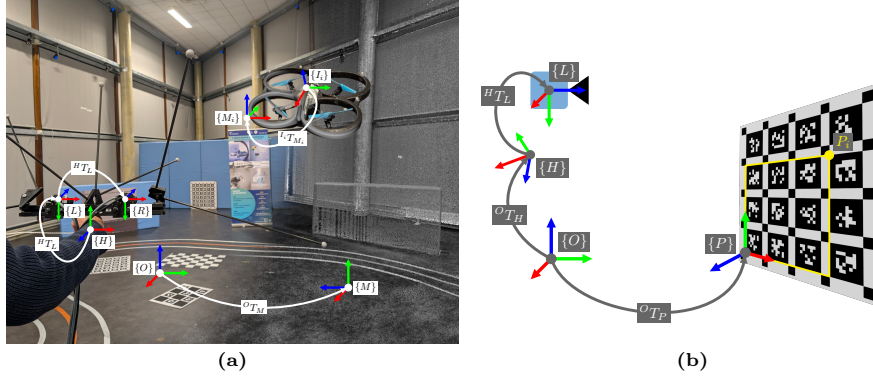


Fig. 3: a) Coordinate frames involved in the reconstruction of the scenes: optitrack (O), scenario model (M), cameras (L , R), camera support/hand (H), i -th IMO (I_i), and model of IMO i (M_i). White arrows represent transformations to be estimated. b) Coordinates frames to compute the position of the AprilGrid points P_i with respect to the left camera frame L .

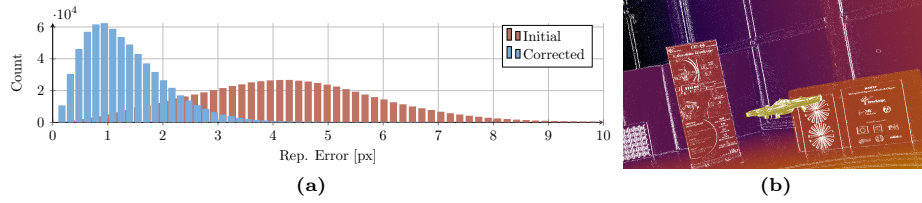


Fig. 4: (a) Histograms of re-projection errors before and after the global refinement of the calibration (cameras + hand-eye). The ground truth projections were obtained with an AprilGrid. (b) Example depth image with superposed events. Events are represented as white dots.

In order to reduce the re-projection errors as much as possible, we iteratively optimized ${}^O T_P$, ${}^H T_L$ and, θ_L by minimizing f using a quasi-Newton algorithm. As shown in Fig. 4a, the optimization considerably improved the errors.

Lastly, to estimate ${}^O T_M$, we take advantage of the rich texture and high-resolution of the scenario model to identify the position of the Optitrack cameras, whose position in O is known. Given the set of corresponding points, the transformation can be easily computed. The average alignment error is of 1.1 cm. A similar process is followed for the transformations between the IMOs' model and their optitrack body ${}^I T_{M_i}$. Conversely, in this case we use the positions of the Optitrack markers, which can be easily located in the objects' model and are known with respect to the optitrack body. As shown in Fig. 4b, this method results in ground truth maps perfectly aligned with the events.

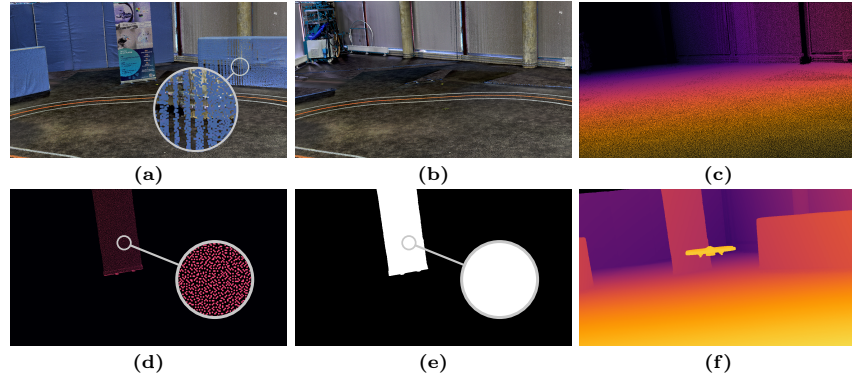


Fig. 5: Process of generation of ground truth depth maps. (a) Static environment model. (b) Environment model without foreground objects. (c) Initial projection of background points. (d) Initial projection of one of the foreground objects. (e) Mask computed for the object in the previous subfigure. (f) Densified depth map with all objects (including one IMO).

5 Ground Truth

5.1 Camera and Object Poses

In order to obtain the pose and velocity of each moving object at each time step of the sequence, optitrack data is filtered and differentiated. Despite the high accuracy of the pose estimates, filtering is required before differentiating to deal with noise coming from the finite resolution of the Optitrack cameras, inaccurate marker detections, and other sources of error. Qualitatively it was found that a simple mean filter with a sliding window of 20 ms was enough to remove the noise. Original and filtered poses are compared in the supplementary material. To estimate velocities, we take 40 ms windows and compute the Lie twists from the first to the last pose of the windows. The linear and angular velocity vectors are obtained from the twist parameters.

5.2 Depth

The depth ground truth is generated from the environment model and the IMO models, placed in the environment given their estimated poses from the optitrack. Then, the resulting point clouds are projected onto the camera image plane, given the camera pose, to obtain a depth image for each of the cameras at the 200 Hz ground truth frequency.

Despite the very high resolution of the models, it is not sufficient to obtain depth maps without pixel gaps. Moreover, as shown in Fig. 5a, background points may be projected onto image areas corresponding to foreground objects. To overcome these problems, we first project the environment model without foreground objects (Fig. 5b) resulting in depth maps as the one in Fig. 5c. Given

the high density of points, maps are completed by assigning the closest depth value to all unlabeled pixels. Then, we separately project all foreground objects (static and dynamic), obtaining a sparse sets of points, as in the example Fig. 5d. To densify objects, a segmentation mask is computed for each object and the same densification strategy is applied inside the mask. Masks are obtained by computing a Delaunay triangulation [20] and filling the triangles with all sides shorter than a threshold of 4 pixels. An example of a resulting mask is shown in Fig. 5e.

The final depth map is obtained by superimposing the foreground maps onto the background map. In case of occlusion between two objects, we take the smaller depth value for each pixel. An example of the results is shown in Fig. 5f. In this process we also generate ground truth masks for object segmentation.

5.3 Optical Flow and Scene Flow

Computing the relative velocity between a point belonging to an object and a camera, all with known pose and velocity, is straightforward. Thus, we compute scene flow ground truth as a 3D velocity vector $[\dot{X} \ \dot{Y} \ \dot{Z}]^T$ for each pixel of the previous depth images with the poses and velocities from subsection 5.1. The equations to compute these vectors are provided in the supplementary material.

Optical flow represents the motion of points on the image plane. If we differentiate the projection function of a point $[X \ Y \ Z]$ with respect to time, we can express optical flow ground truth $[\dot{u} \ \dot{v}]$ as a function of scene flow:

$$\begin{bmatrix} \dot{u} \\ \dot{v} \\ 0 \end{bmatrix} = K \frac{1}{Z^2} \begin{bmatrix} \dot{X}Z - X\dot{Z} \\ \dot{Y}Z - Y\dot{Z} \\ 0 \end{bmatrix} \quad (2)$$

where K is the projection matrix. In this way, ground truth optical flow is calculated for rectified images. By taking into account distortion coefficients, optical flow is also provided for raw camera images.

6 Discussion

Our dataset provides practical event streams and high accuracy ground truth. We put efforts to minimize the remaining minor inaccuracies in calibrations, models and object poses that are unavoidable in all real datasets. While care was taken to properly set the lighting conditions and adjust the camera optics and their biases, the event streams still contain a number of noisy events, which is realistic for event cameras and necessary for correctly benchmarking algorithms in real-world conditions.

As the YCB objects were moved manually, held with a fishing line, some spurious events were also generated by light reflecting off the fishing line. Care was taken to minimize the impact, and often such events were barely distinguished from random noise. Conversely, hands holding the line might be visible in some

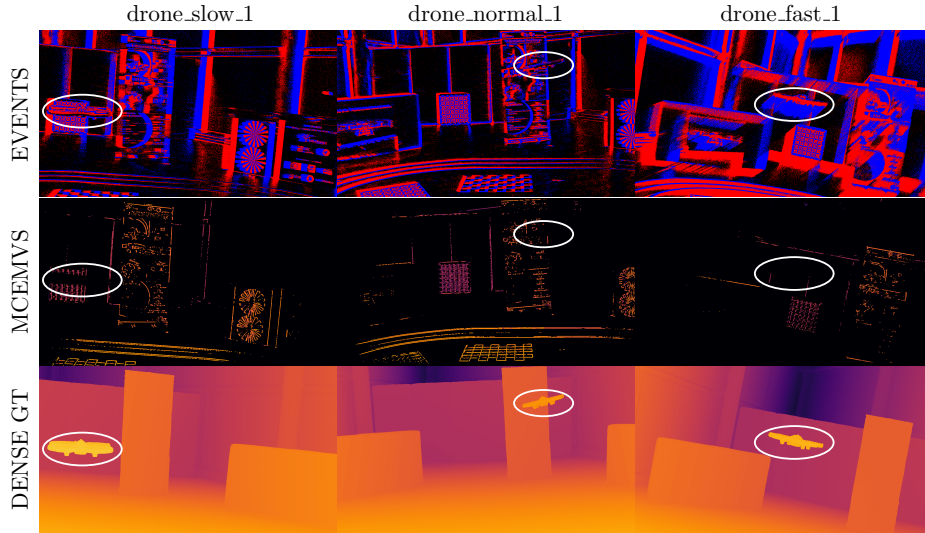


Fig. 6: Depth estimation results in some sequences of E-MOTION. White ellipses indicate the position of the IMO, still challenging to be processed with SOTA methods.

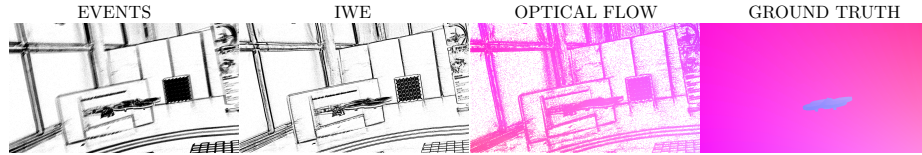


Fig. 7: Optical flow estimation results in the *drone_slow_1* sequence.

sequences during short intervals of time. This actuation mechanism limits the range and type of motion of the YCB objects.

Albeit our dataset lacks an IMU sensor, we provide scripts to simulate it from the camera poses with user-defined noise parameters. Furthermore, we provide scripts to recompute flow ground truth in terms of exact pixel or point displacements instead of relative velocities, which might be preferred when estimating motion for a specific temporal window.

6.1 Data Processing Examples

We showcase the usefulness of our data by estimating depth with MCEMVS [12] and optical flow with contrast maximization [32]. MCEMVS stands out as one of the best long temporal baseline stereo methods for event cameras and compares well with learning models [13]. As qualitatively shown in Fig. 6, MCEMVS is able to provide good depth estimates in the slow and normal camera velocity quadrotor sequences, however, in the fast condition, the number of points for which the depth is calculated decreases considerably. This effectively reflects the

different difficulty levels present in our dataset. Additionally, MCEMVS is based on the assumption of a static scenario and that is why the IMO depth is missing in all sequences. As shown in Fig. 7, the optical flow algorithm also provides inaccurate measurements where the IMO. This highlights the importance of IMOs in the dataset and of methods developed for dynamic scenarios.

7 Conclusion

This work presented E-MOTION, the first large-scale dataset for scene flow estimation recorded with a real stereo setup of event cameras. In addition to scene flow, E-MOTION provides ground truth for other tasks such as depth, optical flow, 6DoF object pose estimation and object segmentation. Another uniqueness of E-MOTION is the high 200 Hz rate of the dense ground truth at high-resolution to fit with the Prophesee EVK4 event cameras. We believe that this dataset opens the door to the development of new event-based methods leveraging stereo high-resolution event information, with precise evaluation of high-frame-rate estimations. Additionally, it serves as a challenging benchmark to evaluate the robustness of algorithms and learning models thanks to its scenarios of varied dynamics.

Acknowledgements

This project is co-funded by the European Union’s Horizon Europe research and innovation programme Cofund SOUND.AI under the Marie Skłodowska-Curie Grant Agreement No. 101081674. Part of this work was funded by the SIVALab joint laboratory between Renault, UTC, and CNRS. The authors thank Professor Eduard Antaluca of the Avenues Lab at UTC for providing the 3D scans of the scenario, and Alejandro Tevera and Alejandro Millan for their assistance with the recordings.

References

1. Event-based vision resources. https://github.com/uzh-rpg/event-based_vision_resources, accessed: 2026-01-19
2. Burner, L., Mitrokhin, A., Fermüller, C., Aloimonos, Y.: Evimo2: an event camera dataset for motion segmentation, optical flow, structure from motion, and visual inertial odometry in indoor scenes with monocular or stereo algorithms. arXiv preprint arXiv:2205.03467 (2022)
3. Calli, B., Singh, A., Bruce, J., Walsman, A., Konolige, K., Srinivasa, S., Abbeel, P., Dollar, A.M.: Yale-cmu-berkeley dataset for robotic manipulation research. *The International Journal of Robotics Research* **36**(3), 261–268 (2017)
4. Chaney, K., Cladera, F., Wang, Z., Bisulco, A., Hsieh, M.A., Korpela, C., Kumar, V., Taylor, C.J., Daniilidis, K.: M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4016–4023 (2023)

5. Dubeau, E., Garon, M., Debaque, B., de Charette, R., Lalonde, J.F.: Rgb-de: Event camera calibration for fast 6-dof object tracking. In: 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 127–135. IEEE (2020)
6. Furgale, P., Rehder, J., Siegwart, R.: Unified temporal and spatial calibration for multi-sensor systems. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1280–1286. IEEE (2013)
7. Furrer, F., Fehr, M., Novkovic, T., Sommer, H., Gilitschenski, I., Siegwart, R.: Evaluation of combined time-offset estimation and hand-eye calibration on robotic datasets. In: Field and Service Robotics: Results of the 11th International Conference. pp. 145–159. Springer (2017)
8. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., et al.: Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* **44**(1), 154–180 (2020)
9. Gao, L., Liang, Y., Yang, J., Wu, S., Wang, C., Chen, J., Kneip, L.: Vector: A versatile event-centric benchmark for multi-sensor slam. *IEEE Robotics and Automation Letters* **7**(3), 8217–8224 (2022)
10. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters* **6**(3), 4947–4954 (2021)
11. Gehrig, M., Millhäusler, M., Gehrig, D., Scaramuzza, D.: E-raft: Dense optical flow from event cameras. In: 2021 International Conference on 3D Vision (3DV). pp. 197–206. IEEE (2021)
12. Ghosh, S., Gallego, G.: Multi-event-camera depth estimation and outlier rejection by refocused events fusion. *Advanced Intelligent Systems* **4**(12), 2200221 (2022)
13. Ghosh, S., Gallego, G.: Event-based stereo depth estimation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
14. Glover, A., Gava, L., Li, Z., Bartolozzi, C.: Edopt: Event-camera 6-dof dynamic object pose tracking. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 18200–18206. IEEE (2024)
15. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., et al.: Kubric: A scalable dataset generator. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3749–3761 (2022)
16. Griwodz, C., Gasparini, S., Calvet, L., Gurdjos, P., Castan, F., Maujean, B., De Lillo, G., Lanthony, Y.: Alicevision meshroom: An open-source 3d reconstruction pipeline. In: Proceedings of the 12th ACM multimedia systems conference. pp. 241–247 (2021)
17. Hay, O.A., Huang, X., Ayyad, A., Sherif, E., Almadhoun, R., Abdulrahman, Y., Seneviratne, L., Abusafieh, A., Zweiri, Y.: E-pose: A large scale event camera dataset for object pose estimation. *Scientific Data* **12**(1), 245 (2025)
18. Hu, Y., Liu, S.C., Delbruck, T.: v2e: From video frames to realistic dvs events. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1312–1321 (2021)
19. Ieng, S.H., Carneiro, J., Benosman, R.B.: Event-based 3d motion flow estimation using 4d spatio temporal subspaces properties. *Frontiers in Neuroscience* **10**, 596 (2017)
20. Lee, D.T., Schachter, B.J.: Two algorithms for constructing a delaunay triangulation. *International Journal of Computer & Information Sciences* **9**(3), 219–242 (1980)

21. Li, Y., Huang, Z., Chen, S., Shi, X., Li, H., Bao, H., Cui, Z., Zhang, G.: Blink-flow: A dataset to push the limits of event-based optical flow estimation. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3881–3888. IEEE (2023)
22. Li, Y., Shen, Y., Huang, Z., Chen, S., Bian, W., Shi, X., Wang, F.Y., Sun, K., Bao, H., Cui, Z., et al.: Blinkvision: A benchmark for optical flow, scene flow and point tracking estimation using rgb frames and events. In: European conference on computer vision. pp. 19–36. Springer (2024)
23. Lin, S., Ma, Y., Guo, Z., Wen, B.: Dvs-voltmeter: Stochastic process-based event simulator for dynamic vision sensors. In: European Conference on Computer Vision. pp. 578–593. Springer (2022)
24. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4040–4048 (2016)
25. Mitrokhin, A., Fermüller, C., Parameshwara, C., Aloimonos, Y.: Event-based moving object detection and tracking. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1–9. IEEE (2018)
26. Mitrokhin, A., Ye, C., Fermüller, C., Aloimonos, Y., Delbruck, T.: Ev-imo: Motion segmentation dataset and learning pipeline for event cameras. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6105–6112. IEEE (2019)
27. Mueggler, E., Rebecq, H., Gallego, G., Delbruck, T., Scaramuzza, D.: The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International journal of robotics research* **36**(2), 142–149 (2017)
28. Olson, E.: Apriltag: A robust and flexible visual fiducial system. In: 2011 IEEE international conference on robotics and automation. pp. 3400–3407. IEEE (2011)
29. Peng, S., Zhou, H., Dong, H., Shi, Z., Liu, H., Duan, Y., Chang, Y., Yan, L.: Cosec: A coaxial stereo event camera dataset for autonomous driving. *arXiv preprint arXiv:2408.08500* (2024)
30. Rebecq, H., Gehrig, D., Scaramuzza, D.: Esim: an open event camera simulator. In: Conference on robot learning. pp. 969–982. PMLR (2018)
31. Rojtblerg, P., Pöllabauer, T.: Ycb-ev 1.1: Event-vision dataset for 6dof object pose estimation. In: European Conference on Computer Vision. pp. 1–13. Springer (2024)
32. Shiba, S., Klose, Y., Aoki, Y., Gallego, G.: Secrets of event-based optical flow, depth and ego-motion estimation by contrast maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(12), 7742–7759 (2024)
33. Sonthalia, A., Batraway, R., Schuster, R., Stricker, D.: Evlidar-flow: Attention-guided fusion between point clouds and events for scene flow estimation. In: ICPRAM. pp. 733–742 (2023)
34. Stoffregen, T., Scheerlinck, C., Scaramuzza, D., Drummond, T., Barnes, N., Kleeman, L., Mahony, R.: Reducing the sim-to-real gap for event cameras. In: European Conference on Computer Vision. pp. 534–549. Springer (2020)
35. Sun, Z., Messikommer, N., Gehrig, D., Scaramuzza, D.: Ess: Learning event-based semantic segmentation from still images. In: European Conference on Computer Vision. pp. 341–357. Springer (2022)
36. Vedula, S., Baker, S., Rander, P., Collins, R., Kanade, T.: Three-dimensional scene flow. In: Proceedings of the Seventh IEEE International Conference on Computer Vision. vol. 2, pp. 722–729. IEEE (1999)

37. Wan, Z., Mao, Y., Zhang, J., Dai, Y.: Rpeflow: Multimodal fusion of rgb-pointcloud-event for joint optical flow and scene flow estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10030–10040 (2023)
38. Wang, Z., Pan, L., Ng, Y., Zhuang, Z., Mahony, R.: Stereo hybrid event-frame (shef) cameras for 3d perception. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9758–9764. IEEE (2021)
39. Zhai, M., Xiang, X., Lv, N., Kong, X.: Optical flow and scene flow estimation: A survey. *Pattern Recognition* **114**, 107861 (2021)
40. Zhou, H., Chang, Y., Shi, Z.: Bring event into rgb and lidar: Hierarchical visual-motion fusion for scene flow. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26477–26486 (2024)
41. Zhu, A., Yuan, L., Chaney, K., Daniilidis, K.: Ev-flownet: Self-supervised optical flow estimation for event-based cameras. In: Proceedings of Robotics: Science and Systems. Pittsburgh, Pennsylvania (June 2018). <https://doi.org/10.15607/RSS.2018.XIV.062>
42. Zhu, A.Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K.: The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters* **3**(3), 2032–2039 (2018)