M3ED: Multi-Robot, Multi-Sensor, Multi-Environment Event Dataset

Kenneth Chaney*†, Fernando Cladera*†, Ziyun Wang†, Anthony Bisulco†,
M. Ani Hsieh†, Christopher Korpela‡, Vijay Kumar†, Camillo J. Taylor†, and Kostas Daniilidis†

Abstract

We present M3ED, the first multi-sensor event camera dataset focused on high-speed dynamic motions in robotics applications. M3ED provides high-quality synchronized and labeled data from multiple platforms, including ground vehicles, legged robots, and aerial robots, operating in challenging conditions such as driving along off-road trails, navigating through dense forests, and performing aggressive flight maneuvers. Our dataset also covers demanding operational scenarios for event cameras, such as scenes with high egomotion and multiple independently moving objects. The sensor suite used to collect M3ED includes high-resolution stereo event cameras (1280×720), grayscale imagers, an RGB imager, a high-quality IMU, a 64-beam LiDAR, and RTK localization. This dataset aims to accelerate the development of event-based algorithms and methods for edge cases encountered by autonomous systems in dynamic environments.

The dataset can be found at https://m3ed.io and the code used to pre-process the data is available at https://github.com/daniilidis-group/m3ed.

1. Introduction

Mobile robotics has increasingly moved towards applications beyond the smooth streets of Karlsruhe, where the pioneering KITTI [11] dataset was obtained. Next-generation robotics perception systems must be able to handle increasingly difficult tasks such as autonomous navigation in rough terrain, aggressive fast motions, and large amounts of mechanical vibration. Event cameras are particularly well suited for these operational scenarios, since they can react and respond with low latencies and high dynamic range [7]. In recent years, event cameras have undergone a dramatic evolution, increasing their resolution and

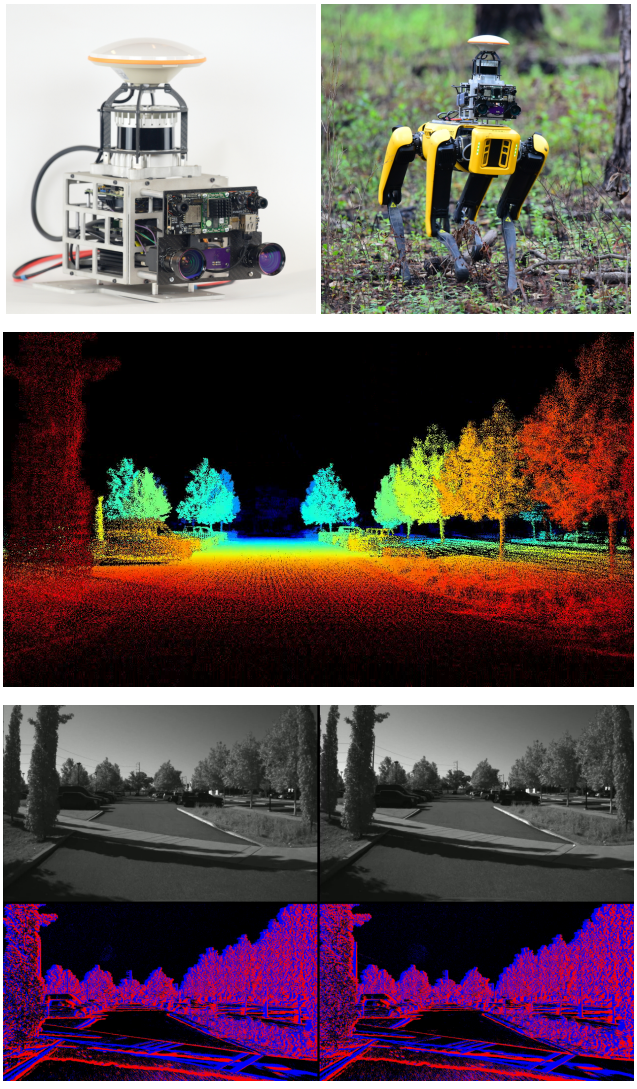

Figure 1. Overview of the methods and results for M3ED: Top left: our modular sensor stack design incorporating all the sensors in a compact package. Top right: Sensor stack on Boston Dynamic Spot robot, one of our three platforms. Middle: Depth estimated using the LiDAR for one of the car sequences. Bottom: Corresponding framed events from Prophesee EVKv4 sensors and grayscale stereo pair.

*These authors contributed equally.
†GRASP Laboratory, University of Pennsylvania, PA, 19104, USA {chaneyk, fclad, ziyunw, abisulco, mya, kumar, cjtaylor, kostas}@seas.upenn.edu
‡Robotics Research Center, Department of Electrical Engineering and Computer Science, United States Military Academy, West Point, NY, 10996, USA. christopher.korpela@westpoint.edu

Dataset	Platform	Terrain	Event Cameras	LiDAR	CIS Cameras	Semantic Labels
The Event Camera Dataset [19]	Slider Hand Held	Urban Indoor	Inivation DVS 240C 240x180	N/A	DVS APS Pixel 240x180 Grayscale	N/A
MVSEC [29]	Car + Motorcycle Quadrotor	Urban Indoor Flight	Inivation DVS 346 346x260	Velodyne VLP-16	Vi-Sensor 752x480 Grayscale	N/A
KITTI 360 [15]	Car	Urban	N/A	Velodyne HDL-64E	YES	37 Classes
DSEC [9]	Car	Urban and Suburban	Prophesee Gen 3 640x480	Velodyne VLP-16	FLIR Backfly S 1440x1080 RGB	11 Classes
VECtor [8]	Helmet + Cart	Indoor	Prophesee Gen 3 640x480	Ouster OS0-128	FLIR Grasshopper3 1224 × 1024 Grayscale	N/A
TUM-VIE [14]	Helmet + handheld	Indoor and Outdoor	Prophesee Gen 4 1280x720	N/A	IDS Camera uEye 1224 × 1024 Grayscale	N/A
M3ED	Car Quadroped UAV	Forest and Urban Forest and Urban Forest and Urban	Prophesee Gen 4 1280x720	Ouster OS1-64U	OVC 3b 1280x800 RGB + Grayscale	11 Classes 3D Instances

Table 1. Comparison of perception datasets acquired with event cameras, CIS cameras and LiDARs, for robotics and automotive applications. M3ED is the only dataset that provides indoor and outdoor sequences using high-resolution stereo event cameras, in heterogeneous robotics platforms.

throughput, making them more attractive compared to conventional CMOS-based image sensors (CISs). However, the widespread adoption of event cameras in robotics is still in its infancy. Event cameras are affected by rapid vibrations and egomotion, which generate a large volume of events, resulting in energy-intensive computation. Furthermore, low-latency segmentation of independently moving objects (IMOs) remains a challenging task for event cameras [5]. We propose an integrated dataset to evaluate event-based algorithms for real-world robotics applications using high-definition (HD) event cameras. Our dataset includes data from a variety of robot platforms:

- A legged robot walking on paved and dirt paths.
- An unmanned Aerial Vehicle (UAV) flying in urban and rural environments, including GPS-denied environments such as under forest canopies.
- A wheeled ground vehicle driving in urban and off-road environments.

Our sensor stack on a legged robot is shown in Fig. 1 along with example data from one of the car sequences. Compared to other publicly available datasets, **M3ED's contributions and novelty include**:

- The first dataset that includes data from high-resolution event cameras mounted on heterogeneous robotic platforms.
- Semantic labeling in unstructured, dynamic environments (off-road, under forest canopies), including LiDAR point clouds, images, and events.
- Ground truth pose, depth, and flow.

In this work, we describe the methods and implementations used for collecting the data, the different platforms and sequences collected, and the future research that this dataset enables.

2. Related Work

Event camera datasets have flourished over the last decade, lowering the barrier of entry for algorithm development. MVSEC [29] pioneered the multi-sensor real-world event camera dataset, providing three primary flight and driving sequences, including night sequences. Ground truth depth was provided by LiDAR, and ground truth pose was derived from a combination of GPS and motion capture data (where applicable).

Other event-based datasets focus on automotive applications: ATIS Automotive dataset [4], N-Cars [24], and 1 MP Detection [20] are event-native datasets with car and pedestrian labels directly on the event sensor. DDD17 [2] and DDD20 [12] recorded car motion information (from the CAN bus) synchronized with event camera streams. Recently, DSEC [9] introduced a KITTI-like dataset providing ground-truth pose and depth with corresponding events. DSEC has been extended for flow [10] and semantics [25].

Beyond automotive applications, event camera datasets have also been published for robotic applications. EvIMO [18] and EvIMO2 [3] study independently moving objects through a combination of high-quality object scans, motion capture, and event-based cameras. These datasets focus on the quality of individual examples with variations in motion. TUM-VIE [14] and VECtor [8] each provide head-mounted ego-centric motion as well as smooth pole-mounted trajectories inside a fixed indoor environment.

Our work aims to provide data sequences for **both automotive and robotic applications**, with **novel platforms** such as legged robots, and **challenging conditions** such as forests and off-road driving. Compared to other datasets in the literature, M3ED provides similar hardware configurations on different platforms, allowing a direct comparison of algorithm performance. Furthermore, our dataset targets heterogeneous conditions beyond driving in urban environments and indoors. Finally, **we release all the raw data**. A summary of the differences between our work and the literature is shown in Table 1.

Sensor Type	Description
OVC 3b	2x 1280x800 Grayscale AR0144 1/4” 12 cm baseline FoV: $61° \times 40°$ 1x 1280x800 RGB AR0144 1/4” FoV: $52° \times 34°$ 1x VectorNav VN100T-SMD
2x Prophesee EVKv4	Prophesee IMX636 1280x720 1/2.5” sensor FoV: $63° \times 38°$
Ouster OS1-64U	64 vertical channels 2048 horizontal points 120 m range 45°vertical FoV
UBlox ZED-F9P	GPS Reciever 5Hz update NTRIP RTK Corrections

Table 2. Sensor stack hardware details. All imagers have similar resolution, field of view, and baseline. RTK accuracy changes greatly depending on environmental conditions.

3. Methods

3.1. Hardware Overview

The individual specifications of each sensor are provided in Table 2. The Open Vision Computer 3b (OVC) [22] provides high-quality hardware-synchronized global shutter stereo images in grayscale, a single RGB image stream, and high-quality IMU measurements. The OVC orchestrates the system and provides synchronization signals to the other sensors. We chose the Prophesee EVKv4 as our main event camera sensors, due to its high resolution (1280×720) and small pixel pitch ($4.86\,\mu m$). The event cameras are placed at an equivalent baseline and a similar field of view to the OVC imagers to provide relevant comparisons for VIO applications. The Ouster OS1-64 provides high-resolution LiDAR for accurate mapping on all platforms. An RTK-GPS module provides RTK GPS when available. Corrections are provided in two manners: an NTRIP Server from a statically calibrated base station when internet is available (urban car and quadruped), or a mobile base station using the U-Blox PointPerfect service, transmitting corrections with a 915 MHz telemetry radio. Finally, all raw data is collected onboard a NUC 10i7FNB.

3.2. Bias and Event Rate Controller

The IMX636ES sensor has a built-in event rate controller (ERC) to limit the number of events received. We evaluated the effectiveness of the ERC and experimentally found that the overall quality of the events was reduced. This ultimately led to the removal of the ERC at the expense of more captured events.

The high data rate can cause issues within the data stream itself. The data path was optimized to avoid unnecessary compute and memory copies. The following optimizations were used: we recorded raw EVT3 packets from Metavision, utilize a single ROS nodelet process for both event cameras, and record the ROS bag itself from within the nodelet process. We confirmed the integrity of the data stream through analysis of the synchronization signals through the camera. The observed error in these signals is approximately 1 microsecond per second, which is well within tolerance of the clocks involved.

Bias tuning provides a way of decreasing the noise events depending on the signal-to-noise ratio (SNR) of the scene. However, we experimentally observed that the number of events in static scenes were greatly modified, even for small changes in the bias. As our goal is to provide comparable data across all sequences, we decided to keep the same bias across all sequences.

3.3. Calibration

Event Cameras Calibration of event cameras was achieved by reconstructing images and calibrating through Kalibr [6]. The image reconstruction was done through the `simple_image_recon` library that is based off the methods described in Frequency Cam [21]. This provides a network-free method for generating images from events that are accurate enough for AprilTag detection.

Lidar Lidar was calibrated by initializing from CAD and refining the rotation for each sequence through geometric alignment of the stereo camera pair.

4. Sequences

The sequences recorded highlight particular challenges for event cameras, such as large egomotion, low light conditions, and noisy environments. Moreover, we targeted scenarios where traditional CIS imagers fail, such as under-canopy in forests or a fast transition from brightness to darkness [7]. For each environment, we provide varying levels of difficulty and lengths. Table 3 provides an overview of the distribution of the sequence types. The remainder of this section provides an insight into why each of the locations was chosen.

4.1. Car Driving

Urban Driving loops were chosen that ranged from dense urban driving to access-controlled parking lots. For instance, *urban_parking*, are easy sequences, where there are no dynamic objects within the scenes, we have constant RTK, and the driving speed is limited. In contrast,

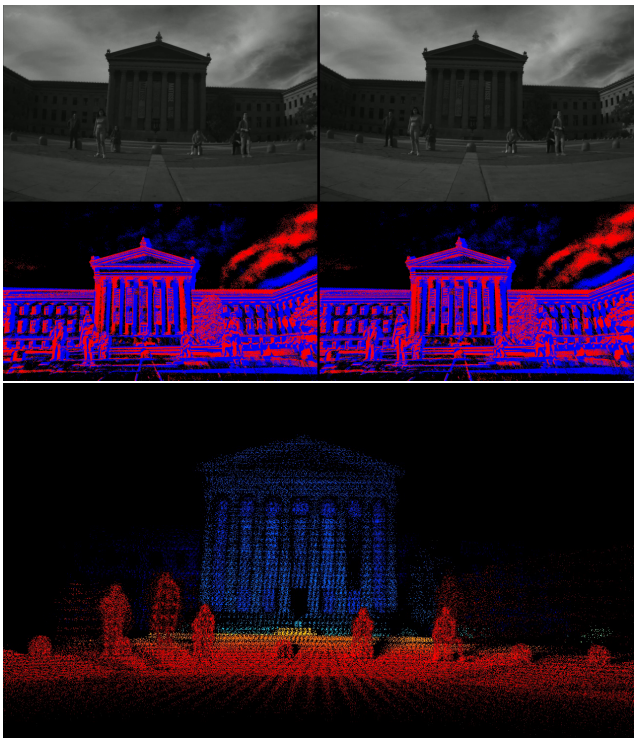

Figure 2. Example data for the *spot_outdoor_steps* sequence showing events with accompanying grayscale and depth.

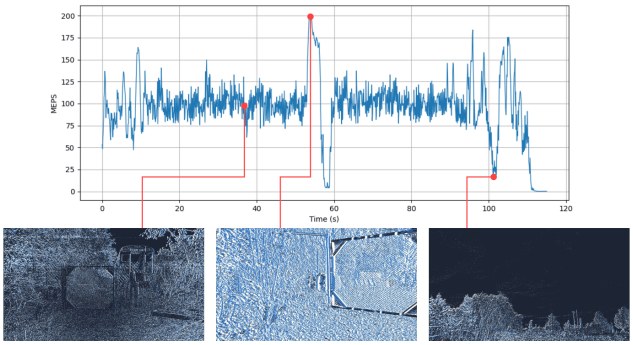

Figure 3. Event cameras for robotic vision face challenges in the number of events that need to be processed. Two leading factors are high motion and high texture. Both of these are exemplified within the *spot_forest* sequences. Left: The average rate during motion is 100 MEPS. Center: Spikes up to 200 MEPS can be seen during rotations of the robot. Right: a portion of the sequence facing up towards the sky, reducing the texture and thus the number of events.

urban_city_hall and *urban_rittenhouse* provide dynamic and highly congested urban scenes of Center City Philadelphia, with traffic and pedestrians, and opportunistic RTK. These sequences were recorded during both day and night conditions, providing contrasting lighting conditions for the same scenes. *urban_schuykill_tunnel* provides a highway tunnel with entrance and exit during the contiguous sequence.

Vehicle	Environment	Total Sequences (test)	Time (s)
Car	Urban	14 (3)	6342
	Forest	3 (1)	485
UAV	Urban	8 (2)	929
	Indoor	3 (1)	171
	Forest	9 (2)	1587
Spot	Urban	11 (2)	1668
	Indoor	3 (1)	287
	Forest	6 (1)	768

Table 3. Total sequences for the dataset. 25% of the sequences will be used as test data. For *car_urban*, we provide day and night sequences. Overall, M3ED provides approximately 3TB of usable data.

Forest These sequences were recorded at the Wharton State Forest, New Jersey. For example, *forest_turnpike_into_ponds* showcases challenging under-canopy images, mud ponds, and dirt roads. *forest_sand* offers an unstable road where the car drifts.

4.2. Legged Robot

This platform is comparatively new within robotics datasets. Legged robots have matured in the past several years and the challenges with perception are starting to emerge. In particular, we see periodic oscillations due to the gait of the robot and sudden jerks when the feet impacts the ground.

Urban The *outdoor_parking* sequences were recorded in an access-controlled parking lot. The *outdoor_skate*, *outdoor_steps*, and *outdoor_under_bridge* sequences show Spot in an open environment with a large number of pedestrians and cyclists (Fig. 2) . These scenes offer a balanced combination of egomotion with a significant number of IMOs.

Indoor The *indoor_obstacles* sequence was recorded at an indoor testing locations for robots. We were able to test Spot's stair locomotion mode, as well as gaits on paved roads. These scenes have few independent moving objects, but particularly high egomotion due to the legged robot gait.

Forest The *forest* sequences have few IMOs but high-texture scenes. These scenes are particularly challenging for algorithms that exploit the sparsity of events to reduce computation [23], as the event count can average 100 MegaEvents per second (MEPS), as shown by Fig. 3.

4.3. UAV

Urban Outdoor flights in a Parking lot, in the *outdoor_parking* sequences, provide a good benchmark for event camera applications in UAVs. There is good RTK coverage, the environment is structured, and the amount of texture and IMOs is low. We provide slow and fast flights in this environment.

Forest Fast autonomous flights in forests have recently attracted more attention in the robotics community, due to the challenges of these cluttered environments [17, 28]. Vision-based perception is particularly difficult due to the high dynamic range required [16]. The *forest* sequences offer opportunities for the development of perception algorithms with event cameras for UAVs in the wild.

5. Dataset Applications

We expect that M3ED will become a new standard for evaluating applications and algorithms of event cameras, such as optical flow estimation, IMO segmentation, and disparity estimation. Approximately 25% of the data will be used for test purposes, and we expect to provide metrics for optical flow, IMO segmentation, disparity estimations, egomotion estimation, and semantics in future works.

5.1. Depth and Pose

FasterLIO [1] provides poses and velocity-corrected LiDAR frames for every sweep of the LiDAR.

Ground truth depth is generated through the projection of accumulated velocity-corrected LiDAR frames provided by FasterLIO. At HD resolutions, the sparseness of the point cloud can be observed when the non-visible portions of the scene show through the projection. This is mitigated by running the hidden point removal (HPR) operator [13] of the point cloud from every viewpoint along the trajectory that is sampled.

Ground-truth pose is also obtained using FasterLIO. This approach enables ground-truth pose for all the scenes with a unified approach, including those scenes where GPS is not available (such as indoors or under-canopy in forests). For our longest sequence between re-observation of a location (800 meters in length), we observed a 2 meter drift over this distance.

5.2. Semantics

The semantics ground truth encompasses 11 categories consistent with the DSEC Semantics extension by Sun et al. [26]. This work utilizes InternImage [27] to generate dense 2D semantic labels. These 11 categories are background, building, fence, person, pole, road, sidewalk, vegetation, car, wall, and traffic sign. An example overlaid frame can be seen in Figure 4. In addition to the 2D semantic labels, we provide 3D instance labels for pedestrians, buildings, cars, and trees. The 3D instance labels are generated using the Segments.ai platform by labeling cuboids for every frame. These IDs are consistent throughout the individual sequences.

5.3. Optical Flow

As an example of current tasks for which M3ED could be used, we give the example of optical flow estimation.

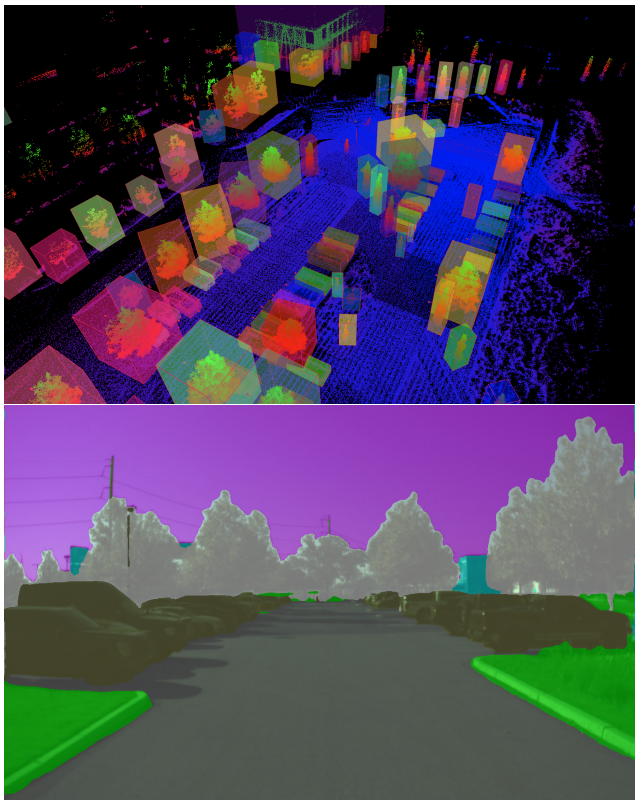

Figure 4. Top: Scene-wide static object instance identification for the *car_urban_parking* data sequence. Bottom: 2D semantic mask generated from InternImage.

Event cameras are naturally suited for optical flow estimation because motion is naturally encoded within the event stream [7]. We evaluated our dataset on the current state-of-the-art optical flow for event cameras, E-RAFT [10]. In Table 4, we show the performance of E-RAFT on selected sequences of our dataset. We use the flow evaluation metrics described in DSEC [9] **NPE**: the percentage of ground-truth pixels with optical flow magnitude error $>$ N. **EPE**: The average L2-Norm of the optical flow error. **AE**: Angular error (degrees).

E-RAFT demonstrated state-of-the-art performance on the DSEC [9] dataset. We used the provided network weights pre-trained on DSEC. We show an example of the flow results of E-RAFT in Fig. 5. To account for the fact that the network has only seen low-resolution events, we also resize the event volumes to 640×480 and then rescale the output flow to 1280×720, marked as E-RAFT. We run the evaluation at the same 10Hz that the network has been trained with. In our experiment, the rapid turns in the sequences cause the network to output erroneous optical flow. This further suggests that a highly dynamic dataset is needed to allow for the further development of event-based perception algorithms.

We would anticipate that the 3D instance labels can be

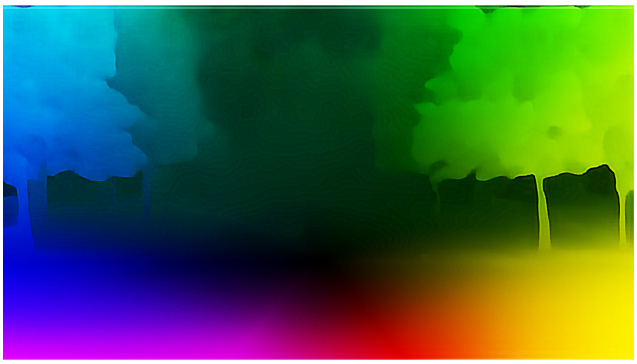

Figure 5. Example flow prediction on the *car_urban_parking* sequence using E-RAFT (color represents direction). The high-resolution events allow rich details to appear in the flow results. Color denotes the direction of the flow at each pixel. Best viewed in color.

Method	EPE	1EP	2EP	3EP	AE
E-RAFT (pre-trained)	5.848	0.934	0.803	0.672	22.886

Table 4. Optical flow performance of E-RAFT [10] on *car_urban_parking* sequence. E-RAFT runs inference on event volumes resized to 640×480 and the flow is scaled back to full resolution.

used to identify IMOs within the scene and act as a mask for future works.

6. Conclusions

6.1. Contributions

This paper described M3ED, a state-of-the-art event-camera dataset generated with heterogeneous sensors. Our goal is that M3ED will allow researchers to better generalize beyond driving or indoor applications, by providing data in challenging conditions that were not explored by other datasets previously.

M3ED also allows the exploration of sensor fusion algorithms in which multiple sensor systems are used to achieve higher levels of robustness. The entire raw training data will be available for download to allow development at any level, as well as the automatic data processing pipeline to generate the output files.

Calibration sequences for camera-to-LiDAR, camera intrinsics and extrinsics, and camera-to-IMU will also be made available to researchers interested in addressing these specific issues.

Overall, we hope that M3ED will become a new standard for event-camera datasets in robotics.

6.2. Limitations

The OVC3b was designed to go on smaller robotic platforms that have size constraints with a baseline of 12 cm. This is a good match for the quadrotor and legged platform, but limits some applications on the driving sequences. Compared to DSEC, the single frame depth estimates will not be as good for this specific scenario.

The lidar and camera field of view overlap only on the top half of the image on the UAV. This limits the usefulness of the lidar for instantaneous understanding of the whole frame. However, the static scenes can still be observed through the generated map. Additionally, the visual sensors are soft mounted relative to the lidar itself and thus the transformation between the lidar and the cameras may shift slightly during sequences (the OVC and the event cameras are rigidly attached). Soft mounting was performed to reduce vibrations on the cameras and IMU.

6.3. Data Availability and Data Content

The data will be fully open and available under **Creative Commons Attribution-ShareAlike 4.0 International**.

We provide time-synchronized HDF5 files for the different sensors of the stack, and simple code snippets to access and process the data. We also provide pre-computed indices to access the closest event timestamp for other sensors.

We also provide HDF5 files with the output of precomputed algorithms used to complement this dataset. For example, as mentioned in Sec. 5.1, provides depth and ground-truth pose. 3D semantic instances Sec. 5.2 obtained through segments.ai are also provided.

Finally, we provide pre-computed extrinsics for all the sensors, as well as intrinsics for the cameras.

Acknowledgements We gratefully acknowledge the support of ARL DCIST CRA W911NF-17-2-0181, ONR grant N00014-20-1-2822, the IoT4Ag Engineering Research Center funded by the National Science Foundation (NSF) under NSF Cooperative Agreement Number EEC-1941529, C-BRIC, a Semiconductor Research Corporation Joint University Microelectronics Program program cosponsored by DARPA, the IARPA ME4AI program, and NSF Award IUCRC 1939132. We gratefully acknowledge Samsung AI 2022-2023 Award to the University of Pennsylvania.

We would like to thank Bernd Pfrommer for his continued support with `metavision_ros_driver` and `simple_image_recon` and Jeremy Wang for support with design and machining of parts.

References

[1] Chunge Bai, Tao Xiao, Yajie Chen, Haoqian Wang, Fang Zhang, and Xiang Gao. Faster-lio: Lightweight tightly coupled lidar-inertial odometry using parallel sparse incremental voxels. *IEEE Robotics and Automation Letters*, 7(2):4861–4868, 2022.

[2] Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. DDD17: End-to-end DAVIS driving dataset. *arXiv preprint arXiv:1711.01458*, 2017.

[3] Levi Burner, Anton Mitrokhin, Cornelia Fermüller, and Yiannis Aloimonos. 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.

[4] Pierre de Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A Large Scale Event-based Detection Dataset for Automotive, 2020.

[5] Davide Falanga, Kevin Kleber, and Davide Scaramuzza. Dynamic obstacle avoidance for quadrotors with event cameras. *Science Robotics*, 5(40):eaaz9712, 2020.

[6] Paul Furgale, Joern Rehder, and Roland Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1280–1286, 2013.

[7] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-Based Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022.

[8] Ling Gao, Yuxuan Liang, Jiaqi Yang, Shaoxun Wu, Chenyu Wang, Jiaben Chen, and Laurent Kneip. VECtor: A Versatile Event-Centric Benchmark for Multi-Sensor SLAM. *IEEE Robotics and Automation Letters*, 7(3):8217–8224, jul 2022.

[9] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021.

[10] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-RAFT: Dense Optical Flow from Event Cameras. In *International Conference on 3D Vision (3DV)*, 2021.

[11] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[12] Yuhuang Hu, Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi 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)*, pages 1–6. IEEE, 2020.

[13] Sagi Katz, Ayellet Tal, and Ronen Basri. Direct visibility of point sets. In *ACM SIGGRAPH 2007 papers*, pages 24–es. 2007.

[14] Simon Klenk, Jason Chui, Nikolaus Demmel, and Daniel Cremers. Tum-vie: The tum stereo visual-inertial event dataset. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8601–8608. IEEE, 2021.

[15] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[16] Xu Liu, Steven W. Chen, Guilherme V. Nardari, Chao Qu, Fernando Cladera Ojeda, Camillo J. Taylor, and Vijay Kumar. Challenges and opportunities for autonomous micro-uavs in precision agriculture. *IEEE Micro*, 42(1):61–68, 2022.

[17] Antonio Loquercio, Elia Kaufmann, René Ranftl, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Learning high-speed flight in the wild. In *Science Robotics*, October 2021.

[18] Anton Mitrokhin, Chengxi Ye, Cornelia Fermüller, Yiannis Aloimonos, and Tobi Delbruck. EV-IMO: Motion segmentation dataset and learning pipeline for event cameras. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6105–6112. IEEE, 2019.

[19] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. 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.

[20] Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to Detect Objects with a 1 Megapixel Event Camera, 2020.

[21] Bernd Pfrommer. Frequency cam: Imaging periodic signals in real-time, 2022.

[22] Morgan Quigley, Kartik Mohta, Shreyas S. Shivakumar, Michael Watterson, Yash Mulgaonkar, Mikael Arguedas, Ke Sun, Sikang Liu, Bernd Pfrommer, Vijay Kumar, and Camillo J. Taylor. The Open Vision Computer: An Integrated Sensing and Compute System for Mobile Robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1834–1840, 2019.

[23] Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. Aegnn: Asynchronous event-based graph neu-

ral networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[24] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. HATS: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1731–1740, 2018.

[25] Zhaoning Sun*, Nico Messikommer*, Daniel Gehrig, and Davide Scaramuzza. ESS: Learning Event-based Semantic Segmentation from Still Images. *European Conference on Computer Vision. (ECCV)*, 2022.

[26] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 341–357. Springer, 2022.

[27] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. *arXiv preprint arXiv:2211.05778*, 2022.

[28] Xin Zhou, Xiangyong Wen, Zhepei Wang, Yuman Gao, Haojia Li, Qianhao Wang, Tiankai Yang, Haojian Lu, Yanjun Cao, Chao Xu, and Fei Gao. Swarm of micro flying robots in the wild. *Science Robotics*, 7(66):eabm5954, 2022.

[29] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018.