PEDRo: an Event-based Dataset for Person Detection in Robotics

Chiara Boretti [†, *] Philippe Bich [†] Fabio Pareschi [†, ¶] Luciano Prono [†]
Riccardo Rovatti [§, ¶] Gianluca Setti [‡, ¶]

[†] DET, [*] PIC4SeR, Politecnico di Torino, Italy
{chiara.boretti, philippe.bich, fabio.pareschi, luciano.prono}@polito.it

[§] DEI, [¶] ARCES, Università di Bologna, Italy
riccardo.rovatti@unibo.it

[‡] CEMSE, King Abdullah University of Science and Technology (KAUST), Saudi Arabia
gianluca.setti@kaust.edu.sa

Abstract

Event-based cameras are devices based on neuromorphic sensors that are gaining popularity in different fields, including robotics. They are suitable for tasks where high-speed, low-latency, low-power operations are required. Person detection is one of these, to allow mobile robots to monitor areas and navigate in crowded environments. Most of the available event-based datasets that contain annotated human figures and collected with a moving camera are designed for autonomous driving tasks. Yet, robotic tasks are certainly not limited to the recognition of pedestrians walking on sidewalks, which makes the above datasets of limited utility. To address this impasse, we introduce a new dataset called PEDRo, which is fully manually labeled. This dataset has been specifically developed for person detection and it counts a total number of 43 259 *bounding boxes included in* 119 *recordings. A moving DAVIS346 event-based camera has been used to collect events in a large variety of indoor and outdoor scenarios with various lighting and meteorological conditions (such as sunny, rainy and snowy). To the best of our knowledge, this is now the largest available dataset for event-based person detection, which has been recorded with a moving camera and manually labeled.*

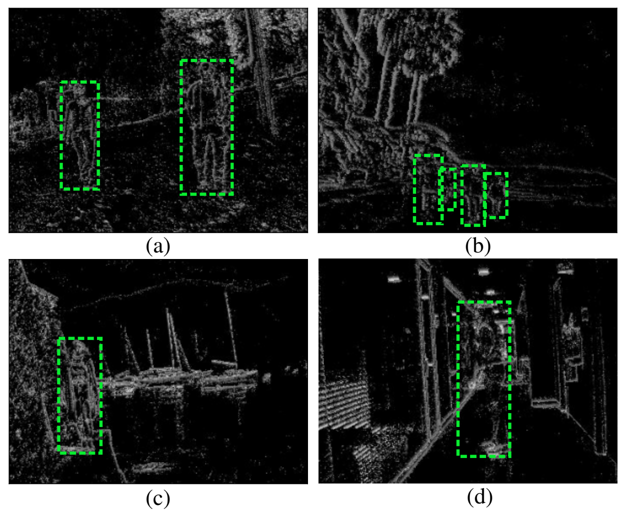

Figure 1. Some of the 43 259 bounding boxes manually annotated contained in PEDRo. The dataset focuses on people and it presents recordings taken in a large variety of environments such as a) woods, b) beaches, c) lakes, d) indoor scenarios.

1. Introduction

Event-based cameras [10, 25] are neuromorphic video recording devices that enable low-power and high temporal-resolution acquisition of visual information. They have gained an increasing popularity in a large variety of fields, including surveillance [37, 38], autonomous driving cars [4, 5, 24, 40] and robotics [2, 8, 41, 43]. The sensors used for these devices offer a fundamentally different way of representing visual information compared to traditional frame-based cameras, providing a sparse and asynchronous events stream where each event corresponds to a change in the scene brightness. More specifically, an event is represented by the coordinates of the pixel location where the illuminance change occurs, as well as by the polarity and the time at which the event has been detected. Event cameras typically provide a very high temporal resolution, with an acquisition time of the order of microseconds and a high dynamic range (up to 120 dB) thanks to the logarithmic characteristic of the pixel response. Moreover, their compact size and low power consumption make them well-suited for deployment in mobile and battery-powered robots. In particular, this kind of sensors has shown great potential for tasks that require fast and accurate motion detection and tracking in a large variety of scenarios with different lighting conditions. As an example, person detection is one of these, to help robots monitoring areas for intruders and navigating through crowded spaces.

The state-of-the-art approach to solve person detection problems is to adopt deep learning architectures, whose proficient training is made possible using large quantities of annotated data. However, the number of event-based datasets is notably smaller compared to what is currently available for frame-based cameras, mainly due to the still limited adoption of event sensors. While synthetic data can be used to tackle this problem, their use can lead to suboptimal performance because of the potential disparities between simulated and real events [11, 39].

In recent years, efforts have been made to create large event-based datasets specifically designed for object detection in the context of autonomous driving [6, 32]. These datasets are recorded using a sensor mounted on the dashboard of a car and data are acquired driving in real-world scenarios. To the best of our knowledge, these are the only datasets containing a large number of labeled people recorded using a moving camera. Because of this, the training of algorithms for person detection in robotics applications becomes a challenge. In fact, robots can move both indoors and outdoors and in a variety of environments which are not limited to roads or highways. Moreover, pedestrians are typically walking on the sidewalks far from the camera, while when dealing with robots navigation, people can be close and their figure may be partially out of frame.

To overcome this, we introduce in this work an event-based dataset for Person Detection in Robotics applications (PEDRo), made of 119 recordings of variable duration taken with a moving DAVIS346 camera. The dataset, containing 43 259 bounding boxes, was manually annotated and focuses on human figures in a large variety of environments, atmospheric and lighting conditions, as in Fig. 1. To the best of our knowledge, this is also the largest manually annotated event-based dataset for person detection recorded with a moving camera.

2. Related Works

Event-based datasets Early examples of event-based datasets have been generated from already existing frame-based datasets by extracting events from frames. In [31], the MNIST [23] and the Caltech-101 [9] datasets have been converted into events stream by moving an event camera in front of a screen where the frames were displayed. A similar strategy has been adopted in [18], where the authors generate events datasets from image datasets [13, 22, 36], and where the event-based camera was still and the motion was performed by sliding the images on the monitor. These approaches allow to obtain very large datasets without the need for manual labeling, but the 2D characteristic of the screen and its limited refresh rate negatively impact the quality of the events stream. An alternative procedure has been proposed in [11] where Gehrig *et al*. leverage the capabilities of event simulators, like [19, 34], to transform popular frame-based datasets into their event-based counterparts. As such, clearly real data are still needed to fully exploit the properties of the event cameras and to accurately replicate noise and sensor nonidealities.

Over the past few years, the number of event datasets collected by using event-based sensors has in fact increased. For example, in [1, 16] and in [44] two datasets collected in driving scenarios with a single and a stereo configuration of a DAVIS346 camera respectively are presented to estimate quantities such as the steering angle, depth, and odometry.

Event-based datasets for person detection More recently, the number of event-based datasets for person detection has increased. In [27], Shu *et al*. present a small dataset completely focused on pedestrians composed of 12 sequences recorded using a fixed DAVIS346 event camera containing 4670 labeled people. This dataset is mainly used for surveillance and it contains also sequences for action recognition and fall detection. Another example is the one produced by Bolten *et al*. in [3] collected with three fixed CeleX-4 DVS [14] event cameras. It features recordings of an outdoor urban public area, and it has been developed for long time monitoring purposes. In 2020, Prophesee released two large datasets recorded in driving scenarios which are the GEN1 Automotive Detection Dataset [6] and the 1 Megapixel Automotive Detection Dataset [32]. The former presents a total of 255 781 manually-labeled bounding boxes (228 123 cars, 27 658 pedestrians) collected for 39 hours with a Prophesee Gen1 sensor [33] at a resolution of 304×240 pixels, while the latter is the most comprehensive event-based dataset to date and it counts 15 hours of recordings with a resolution of 1280×720 pixels and 25 million automatically-labeled bounding boxes. The people labeled in these two datasets are mainly pedestrians walking on sidewalks.

Event-based cameras in Robotics The interest in event-based cameras is increasing also in robotics and their usage has been investigated for Simulatenous Localization And Mapping (SLAM) [35], control strategies [2, 43], and the estimation of quantities like the pose [29, 46], the optical flow [12, 42, 45] and the trajectory of moving objects [15, 28]. Event-based sensors find applications in diverse tasks, ranging from guiding ground robots [2] or underwater platforms [43] to developing obstacle avoidance or SLAM algorithms for Unmanned Aerial Vehicles (UAVs) [8, 41].

3. Dataset

PEDRo is a dataset recorded with a moving camera and completely focused on person detection, specifically designed for service robotics applications[1] It contains 43 259 bounding boxes associated to 27 000 samples with a duration of 40 ms each, extracted from 119 recordings with an average duration of 18 s. Each sample features at least one and at most six bounding boxes. The dataset presents a wide variety of indoor and outdoor scenarios, ranging from office and house environments to mountains, lakes landscapes and seafronts. The recordings are taken with different meteorological conditions such as sunny, rainy and snowy during day and night. The dataset has been collected in 6 months from September 2022 to February 2023 and the people recorded in PEDRo range from 20 to 70 years of age. Most of the labeled subjects are walking, although there are examples of people standing still, sitting, or running. We obtained informed written consent from all the recorded individuals, and we further protect their privacy by publishing only the events and the labels of the recordings.

3.1. Data collection and labeling

The dataset has been entirely recorded using a DAVIS346 event-based camera [20] which has a resolution of 346×260 pixels and outputs simultaneously events and grayscale frames. The camera is in motion, it has been hand-carried to capture the events and the position of the sensor varies among recordings. The dataset has been fully manually labeled by the authors using the grayscale frames and all the bounding boxes have been double-checked. In our case, automatic labeling performed by state-of-the-art object detection models like YOLOv8x [21] does not provide reliable results, probably due to the low resolution of the grayscale frames offered by the camera.

In order to highlight this, we evaluate the quality of the predicted bounding boxes using the Intersection over Union (IoU) as a criterion. More specifically, given two bounding boxes, one true (manually labeled) and one predicted, the IoU measures their degree of overlap as the ratio of their intersection area to their union area (*e.g.*, 1 means perfect overlap, 0 no overlap). Since a frame can contain multiple bounding boxes, we select the optimal true-predicted couples and then we evaluate the average IoU. The optimal true-predicted couples are selected as follows:

- the IoU is evaluated for all the combinations of true and predicted bounding boxes;
- the true and predicted bounding boxes associated with the maximum IoU value are selected as an optimal couple and removed from the computation, and this is repeated until there are no couples remaining;
- all the unpaired bounding boxes count as zero IoU value.

[1]PEDRo is publicly available and it can be obtained via the following link: https://github.com/SSIGPRO/PEDRo-Event-Based-Dataset.git

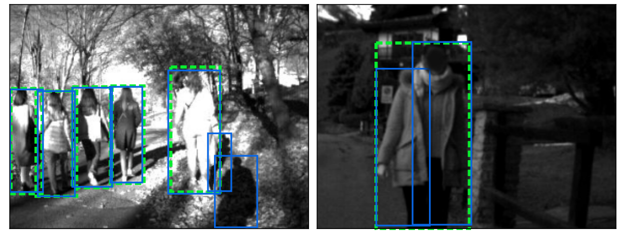

Figure 2. Bounding boxes predicted by the YOLOv8x model are shown with solid light blue lines, while the manually annotated labels are shown with dashed green lines.

With the largest pretrained YOLOv8 model, 22% of the 27 000 greyscale frames used for labeling do not reach an average IoU of 0.85 and almost 45% do not reach an average IoU of 0.90. This means that, for this dataset, automatic labeling is not a suitable option to obtain high-quality ground truth. Fig. 2 shows two frames that result to be wrongly labeled using the YOLOv8x model.

3.2. Dataset Format

The 119 recordings that compose the dataset have been split into train, validation and test subsets. To avoid overlap of data, every single recording belongs entirely to one of these three groups. The number of bounding boxes in the dataset is 43 259, of which 34 243 (79.2%) in train, 4372 (10.1%) in validation and 4179 (9.7%) in test.

Each subset is associated to a text file that lists the recordings and the names of the samples it contains. With sample, we refer to the stream of events collected in a time interval of 40 ms preceding the timestamp of the frame used to obtain the corresponding labels. This time interval is determined by the acquisition rate of the grey-scale images used for the manual labeling process (25 fps).

The events (with positive and negative polarity) are stored in a numpy structure, while their corresponding labels are provided in Pascal VOC format [7]. Each sample can be coupled with its corresponding label by considering the matching names (*e.g.* file `frame0000001.npy` is associated with `frame0000001.xml`).

3.3. Analysis and Statistics

To better highlight some peculiarities of our dataset, we extract some statistics and we compare them to the ones obtained from the GEN1 Automotive Detection Dataset [6]. We select the GEN1 dataset as it is the most extensive automotive dataset with hand-labeled annotations and its spatial resolution is similar to PEDRo.

We start this analysis by looking at the distribution of the bounding boxes in the two datasets using heatmaps, com-

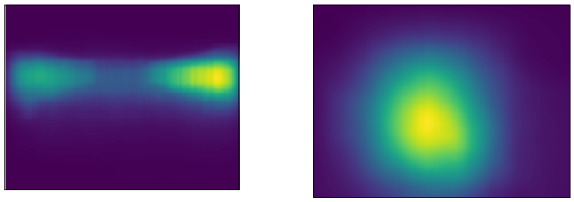

Figure 3. On the left, the heatmap displaying labeled bounding boxes for pedestrians of the GEN1 dataset, while on the right, the heatmap for people in our dataset.

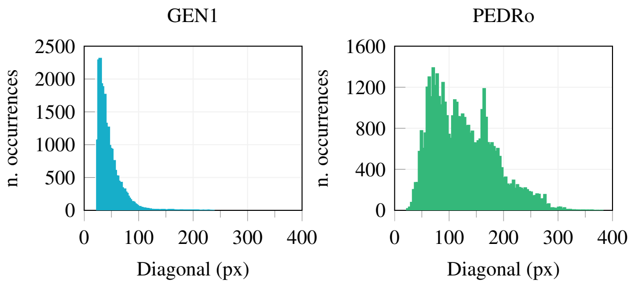

Figure 4. On the left, the distribution of the length (in pixels) of the diagonals of bounding boxes for pedestrians in the GEN1 dataset, while on the right, the diagonals for people in our dataset.

puted by counting the number of bounding boxes covering each pixel over the entire dataset. We normalize the count by dividing it by the total number of bounding boxes in the datasets, resulting in heatmaps containing values in the range $[0, 1]$ as in Fig. 3. Since the GEN1 dataset is collected using a camera on a vehicle, detected people are typically pedestrians walking on sidewalks, so bounding boxes are confined to the image margins. On the contrary, our dataset focuses more on people which are closer to the center of the image and to the camera. Moreover, Fig. 4 shows the distribution of sizes of bounding boxes diagonals in the GEN1 and in our dataset. The automotive dataset presents a long tail distribution and most of the bounding boxes have a small diagonal while our dataset features more bounding boxes with a wider range of diagonal sizes. These characteristics highlight how PEDRo features labeled data whose properties are missing in automotive datasets like GEN1.

4. Experimental Results

In this section we evaluate the performance of YOLOv8x on PEDRo and we compare it with the performance on GEN1. For each dataset, we start from the pre-trained YOLOv8x architecture and we train the model for 5 epochs with Stochastic Gradient Descent (SGD), learning rate of 0.01 and a batch size of 64. Each input sample consists in events organized in a Surface of Active Events (SAE) [27, 30]. With SAE, the events from a sample are aggregated together in a single two-channel tensor (*i.e.* a two-channel image). One channel is generated from positive events, while the other from negative ones. For each channel, the value corresponding to each pixel is proportional to the timestamp of the most recent event, *i.e.* the more recent is the event, the higher is the value. In order to make SAE compliant with YOLOv8x, the values are normalized between 0 and 255 (which is typical for images). Furthermore, since the pre-trained YOLOv8x model accepts 3 channels images, we have removed the third input channel from the model.

Test \ Train	GEN1	PEDRo	GEN1 + PEDRo
GEN1	0.716|0.341	0.487|0.205	0.718|0.342
PEDRo	0.437|0.237	0.895|0.586	0.794|0.504

Table 1. Results expressed as $mAP_{50}|mAP_{50:95}$ with YOLOv8x trained on different training sets and evaluated on various test sets.

To compare the performance, we consider the mean Average Precision value (mAP) [26] which is a widely used metric for understanding the performances of a neural network employed for an object detection task [17, 32]. Tab. 1 presents the results in terms of mAP_{50} and $mAP_{50:95}$ obtained on different test sets. Here we see that GEN1 and PEDRo actually are uncorrelated datasets, being the mAP values quite low when cross-training YOLOv8x, *i.e.* when a model trained on GEN1 is tested on PEDRo and vice-versa. This means that GEN1, being an automotive dataset focused on pedestrians, is not suited for person detection where human figures are the main subject of the scenes. On the contrary, PEDRo allows the training of structures capable of detecting persons closer to the camera and in different environments, which is relevant for different robotic and surveillance applications. Furthermore, by training YOLOv8x with GEN1 in conjunction with PEDRo, there is no loss in performance on the recognition of pedestrians while the model becomes also capable of recognizing targets contained in PEDRo.

5. Conclusions

This paper presents PEDRo which is, to the best of our knowledge, the largest manually annotated event-based dataset recorded with a moving camera and explicitly designed for people detection. This dataset focuses on people, with a wide variety of recording environments and different lighting conditions, making it a relevant addition to other existing event-based datasets, such as the GEN1 Automotive dataset. PEDRo presents an opportunity for enhancing the prediction capabilities of deep learning object detection models with events as input, which can pave the way for novel research avenues and potential applications in diverse fields including but not limited to robotics and surveillance.

References

[1] Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. DDD17: End-To-End DAVIS Driving Dataset. *arXiv*, 2017. 2

[2] Hermann Blum, Alexander Dietmüller, Moritz Milde, Jörg Conradt, Giacomo Indiveri, and Yulia Sandamirskaya. A neuromorphic controller for a robotic vehicle equipped with a dynamic vision sensor. In *Robotics Science and Systems (RSS)*, July 2017. 1, 2

[3] Tobias Bolten, Regina Pohle-Frohlich, and Klaus D. Tonnies. DVS-OUTLAB: A Neuromorphic Event-Based Long Time Monitoring Dataset for Real-World Outdoor Scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1348–1357, June 2021. 2

[4] Guang Chen, Hu Cao, JoÖrg Conradt, Huajin Tang, Florian Rohrbein, and Alois Knoll. Event-Based Neuromorphic Vision for Autonomous Driving: A Paradigm Shift for Bio-Inspired Visual Sensing and Perception. *IEEE Signal Processing Magazine*, 37(4):34–49, 2020. 1

[5] Guang Chen, Fa Wang, Weijun Li, Lin Hong, Jörg Conradt, Jieneng Chen, Zhenyan Zhang, Yiwen Lu, and Alois Knoll. NeuroIV: Neuromorphic Vision Meets Intelligent Vehicle Towards Safe Driving With a New Database and Baseline Evaluations. *IEEE Transactions on Intelligent Transportation Systems*, 23(2):1171–1183, 2022. 1

[6] Pierre de Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A Large Scale Event-based Detection Dataset for Automotive. *arXiv*, Jan. 2020. 2, 3

[7] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vision*, 88(2):303–338, June 2010. 3

[8] Davide Falanga, Kevin Kleber, and Davide Scaramuzza. Dynamic obstacle avoidance for quadrotors with event cameras. *Science Robotics*, 5(40):1–15, 2020. 1, 2

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):594–611, Apr. 2006. 2

[10] 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 Trans. Pattern Anal. Mach. Intell.*, 44(1):154–180, Jan. 2022. 1

[11] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to Events: Recycling Video Datasets for Event Cameras. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3583–3592, June 2020. 2

[12] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-RAFT: Dense Optical Flow from Event Cameras. In *International Conference on 3D Vision (3DV)*, pages 197–206, Dec. 2021. 2

[13] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 Object Category Dataset. Technical Report 7694, California Institute of Technology, 2007. 2

[14] Menghan Guo, Jing Huang, and Shoushun Chen. Live demonstration: A 768 × 640 pixels 200Meps dynamic vision sensor. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–1, May 2017. 2

[15] Botao He, Haojia Li, Siyuan Wu, Dong Wang, Zhiwei Zhang, Qianli Dong, Chao Xu, and Fei Gao. FAST-Dynamic-Vision: Detection and Tracking Dynamic Objects with Event and Depth Sensing. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3071–3078, Sept. 2021. 2

[16] 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 *IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6, Sept. 2020. 2

[17] Yuhuang Hu, Tobi Delbruck, and Shih-Chii Liu. Learning to Exploit Multiple Vision Modalities by Using Grafted Networks. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 85–101, Aug. 2020. 4

[18] Yuhuang Hu, Hongjie Liu, Michael Pfeiffer, and Tobi Delbruck. DVS Benchmark Datasets for Object Tracking, Action Recognition, and Object Recognition. *Frontiers in Neuroscience*, 10, 2016. 2

[19] Yuhuang Hu, Shih-Chii Liu Liu, and Tobi Delbruck. v2e: From Video Frames to Realistic DVS Events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1312–1321, June 2021. 2

[20] IniVation. *DAVIS346 datasheet*, Aug. 2019. 3

[21] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, 1 2023. 3

[22] Matej Kristan, Jiri Matas, Ale Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Toma Vojir, Gustav Hager, Georg Nebehay, Roman Pflugfelder, Abhinav Gupta, Adel Bibi, Alan Lukezic, Alvaro Garcia-Martin, Amir Saffari, Alfredo Petrosino, and Andrés Solis Montero. The Visual Object Tracking VOT2015 Challenge Results. In *IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 564–586, Dec. 2015. 2

[23] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov. 1998. 2

[24] Jianing Li, Siwei Dong, Zhaofei Yu, Yonghong Tian, and Tiejun Huang. Event-Based Vision Enhanced: A Joint Detection Framework in Autonomous Driving. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1396–1401, July 2019. 1

[25] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128 X 128 120db 30mw asynchronous vision sensor that responds to relative intensity change. In *IEEE International Solid State Circuits Conference - Digest of Technical Papers*, pages 2060–2069, Feb. 2006. 1

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 740–755, Sept. 2014. 4

[27] Shu Miao, Guang Chen, Xiangyu Ning, Yang Zi, Kejia Ren, Zhenshan Bing, and Alois Knoll. Neuromorphic Vision Datasets for Pedestrian Detection, Action Recognition, and Fall Detection. *Frontiers in Neurorobotics*, 13, 2019. 2, 4

[28] Anton Mitrokhin, Cornelia Fermüller, Chethan Parameshwara, and Yiannis Aloimonos. Event-Based Moving Object Detection and Tracking. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–9, Oct. 2018. 2

[29] Elias Mueggler, Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. Continuous-Time Visual-Inertial Odometry for Event Cameras. *IEEE Transactions on Robotics*, 34(6):1425–1440, 2018. 2

[30] 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. 4

[31] Garrick Orchard, Ajinkya Jayawant, Gregory K. Cohen, and Nitish Thakor. Converting Static Image Datasets to Spiking Neuromorphic Datasets Using Saccades. *Frontiers in Neuroscience*, 9, 2015. 2

[32] Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to Detect Objects with a 1 Megapixel Event Camera. In *Adv. Neural Inform. Process. Syst.*, volume 33, pages 16639–16652, 2020. 2, 4

[33] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. A QVGA 143 dB Dynamic Range Frame-Free PWM Image Sensor With Lossless Pixel-Level Video Compression and Time-Domain CDS. *IEEE Journal of Solid-State Circuits*, 46(1):259–275, 2011. 2

[34] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an Open Event Camera Simulator. In *Conference on Robot Learning (CoRL)*, pages 969–982, Oct. 2018. 2

[35] Henri Rebecq, Timo Horstschaefer, Guillermo Gallego, and Davide Scaramuzza. EVO: A Geometric Approach to Event-Based 6-DOF Parallel Tracking and Mapping in Real Time. *IEEE Robotics and Automation Letters*, 2(2):593–600, Apr. 2017. 2

[36] Kishore K. Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, July 2013. 2

[37] Juan Pablo Rodríguez-Gomez, A. Gómez Eguíluz, José Ramiro Martínez-de Dios, and Anibal Ollero. Asynchronous event-based clustering and tracking for intrusion monitoring in UAS. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 8518–8524, May 2020. 1

[38] Juan Pablo Rodríguez-Gómez, Augusto Gómez Eguíluz, José Ramiro Martínez-De Dios, and Anibal Ollero. Auto-Tuned Event-Based Perception Scheme for Intrusion Monitoring With UAS. *IEEE Access*, 9:44840–44854, 2021. 1

[39] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the Sim-to-Real Gap for Event Cameras. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 534–549, Nov. 2020. 2

[40] Abhishek Tomy, Anshul Paigwar, Khushdeep S. Mann, Alessandro Renzaglia, and Christian Laugier. Fusing Event-based and RGB camera for Robust Object Detection in Adverse Conditions. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 933–939, May 2022. 1

[41] Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Ultimate SLAM? Combining Events, Images, and IMU for Robust Visual SLAM in HDR and High-Speed Scenarios. *IEEE Robotics and Automation Letters*, 3(2):994–1001, Apr. 2018. 1, 2

[42] Chengxi Ye, Anton Mitrokhin, Cornelia Fermüller, James A. Yorke, and Yiannis Aloimonos. Unsupervised Learning of Dense Optical Flow, Depth and Egomotion with Event-Based Sensors. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5831–5838, 2020. 2

[43] Ibrahim Youssef, Mehmet Mutlu, Behzad Bayat, Alessandro Crespi, Simon Hauser, Jörg Conradt, Alexandre Bernardino, and Auke Ijspeert. A Neuro-Inspired Computational Model for a Visually Guided Robotic Lamprey Using Frame and Event Based Cameras. *IEEE Robotics and Automation Letters*, 5(2):2395–2402, Apr. 2020. 1, 2

[44] 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, July 2018. 2

[45] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras. In *Robotics Science and Systems (RSS)*, June 2018. 2

[46] Yi–Fan Zuo, Jiaqi Yang, Jiaben Chen, Xia Wang, Yifu Wang, and Laurent Kneip. DEVO: Depth-Event Camera Visual Odometry in Challenging Conditions. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2179–2185, May 2022. 2