

# DSEC: A Stereo Event Camera Dataset for Driving Scenarios

Mathias Gehrig, Willem Aarents, Daniel Gehrig and Davide Scaramuzza

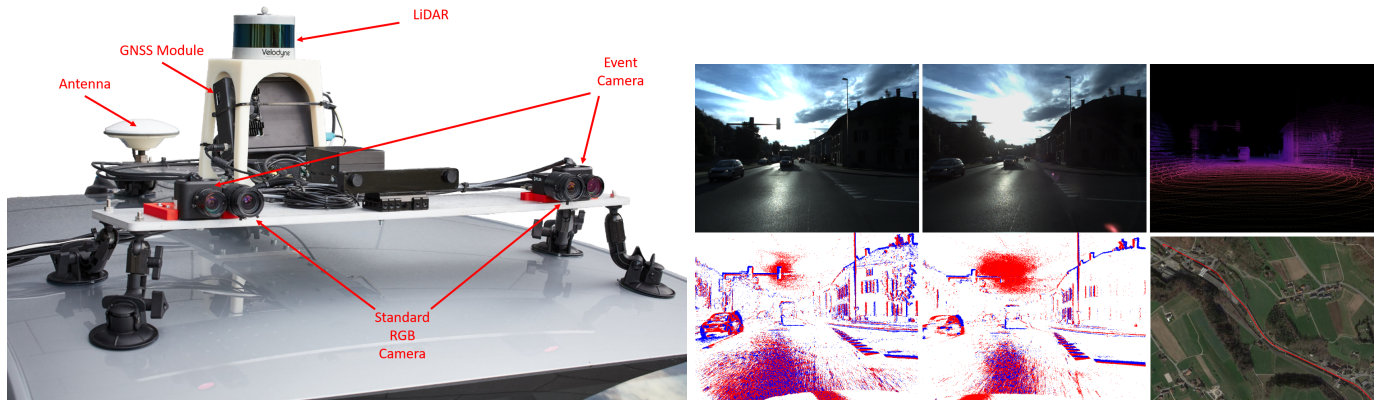


Fig. 3: Left: The sensor suite mounted on top of a car, ready for data collection. Right: Stereo RGB and event camera facing direct sunlight. The dataset contains depth maps for stereo matching and RTK GPS for accurate position measurement.

**Abstract**—Once an academic venture, autonomous driving has received unparalleled corporate funding in the last decade. Still, operating conditions of current autonomous cars are mostly restricted to ideal scenarios. This means that driving in challenging illumination conditions such as night, sunrise, and sunset remains an open problem. In these cases, standard cameras are being pushed to their limits in terms of low light and high dynamic range performance. To address these challenges, we propose, DSEC, a new dataset that contains such demanding illumination conditions and provides a rich set of sensory data. DSEC offers data from a wide-baseline stereo setup of two color frame cameras and two high-resolution monochrome event cameras. In addition, we collect lidar data and RTK GPS measurements, both hardware synchronized with all camera data. One of the distinctive features of this dataset is the inclusion of high-resolution event cameras. Event cameras have received increasing attention for their high temporal resolution and high dynamic range performance. However, due to their novelty, event camera datasets in driving scenarios are rare. This work presents the first high resolution, large scale stereo dataset with event cameras. The dataset contains 53 sequences collected by driving in a variety of illumination conditions and provides ground truth disparity for the development and evaluation of event-based stereo algorithms.

**Index Terms**—Data Sets for Robotic Vision; Data Sets for Robot Learning; Computer Vision for Transportation

Code and dataset: <https://dsec.ifi.uzh.ch>

Manuscript received: October 15th, 2020; Revised January 19th, 2021; Accepted February 16th, 2021.

This paper was recommended for publication by Editor Cesar Cadena Lerma upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the Swiss National Science Foundation (SNSF) through the National Center of Competence in Research (NCCR) Robotics and the SNSF-ERC Starting Grant.

The authors are with the Robotics and Perception Group, Dep. of Informatics, University of Zurich, and Dep. of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland (<http://rpg.ifi.uzh.ch>).

Digital Object Identifier: 10.1109/LRA.2021.3068942

## I. INTRODUCTION

**A**UTONOMOUS driving has received increasing attention in the robotics and computer vision communities over the past decade. However, the robust performance of vision systems in challenging conditions is still an open problem.

Event cameras show promise to alleviate these challenges by providing complementary sensory information to standard cameras. In contrast to standard cameras that capture frames at regular intervals, event cameras report per-pixel brightness changes as a stream of asynchronous events. Event cameras offer significant advantages compared to standard cameras, such as high temporal resolution (microseconds), high dynamic range ( $> 120\text{dB}$ ), less motion blur, and sub-millisecond latency, all of which are crucial for autonomous driving and automotive scenarios. For a detailed survey about event cameras, we refer to [1].

While event camera research has gained momentum in the last few years, event camera datasets in driving scenarios are still a rarity. This can be attributed to the fact that the sensor has only recently been commercialized and thus does not yet enjoy the same availability of standard cameras. The shortage of large scale datasets impedes the event camera research community because it restricts the evaluation to small scale environments that can quickly saturate the performance or generalizability of proposed algorithms.

This work specifically addresses the lack of large scale stereo datasets. To the best of our knowledge, MVSEC [2] is the only dataset with a stereo event camera in driving scenarios. A significant drawback of MVSEC for stereo algorithms is the small resolution of the DAVIS346 cameras (0.1 Megapixels) and the small baseline of 10 centimeters on

top of that. This is in stark contrast to state-of-the-art standard benchmarks such as KITTI [3], [4], [5] or DrivingStereo [6] with camera baselines of approximately 55 centimeters and camera resolution over 1 Megapixel (MP). Both aspects are crucial for stereo algorithms as it is well known that the depth error is approximately inversely proportional to the product of the baseline and focal length [7].

While event cameras certainly possess unique assets compared to standard cameras, estimating dense depth using only stereo event cameras is an ill-posed problem. In a static scene with constant illumination, a static event camera will not trigger any events and can therefore not perform stereo matching. In the case of a dynamic scene and a static camera, it is at least possible to estimate the depth of the dynamic part of the scene. Still, the majority of event-based stereo methods only estimate sparse or semi-dense depth [8], [9], [10], [11]. We refer to the survey by Steffen et al. for an overview of neuromorphic stereo vision approaches [12]. For this reason, we also equip our data collection setup with two high-resolution standard cameras that can allow for dense depth estimation even in cases of static scenes with no motion.

The fusion of frames and events has proven to be beneficial in several prior works [13], [14], [15], [16], [17]. Nevertheless, the underlying assumption of all these approaches is that frames and events share the same pixels. This is the case for example for the older DAVIS cameras [18] but not for state-of-the-art event cameras [19]. Our data collection setup, depicted in figure 3, separates standard and event cameras spatially such that each stereo camera pair is mounted next to each other. By doing so, we minimize the parallax effect between event and standard cameras and facilitate sensor fusion of both modalities. Hence, the proposed dataset raises novel challenges for sensor fusion algorithms which can take advantage of high resolution event and standard cameras.

Finally, we collect data in both favorable and very challenging illumination conditions such as depicted in fig. 3. For real-world applicability, computer vision systems must show robustness also in the most adverse conditions such as driving at night, facing direct sunlight or fast moving objects. This dataset incorporates multiple sequences with precarious illumination conditions that challenge standard and event cameras. We believe that this dataset serves as a foundation to better understand the performance of novel algorithms in various illumination conditions and challenging situations.

In summary, we propose a new dataset that

- addresses the need for a large scale outdoor stereo event camera datasets
- offers the novel challenge of combining events and frames without the common assumption of accurate pixel correspondence.
- provides sequences with various illumination conditions, from favorable to very challenging ones.

## II. RELATED WORK

A multitude of driving datasets for computer tasks has been released in the last decade. This section serves as a brief overview of datasets in the driving domain.

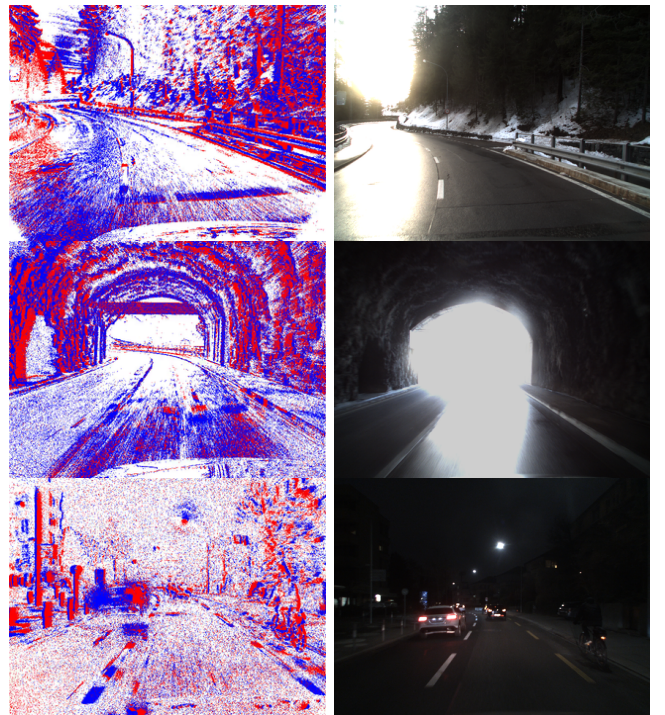


Fig. 3: Examples of challenging illumination conditions from the recordings. The first two examples show that the frame-based camera is prone to saturation of pixels due to the lower dynamic range. The third row is a driving scene at night in which the bicyclist on the right side of the street is difficult to spot.

Following the seminal work of Geiger et al. [3], multiple datasets have been published mostly for more specialized purposes. KITTI 2015 [4] extended the original KITTI version to provide scene flow ground truth. Later KITTI 360 [5] proposed a novel labeling scheme and data for 2D and 3D semantic segmentation. Similarly, Cityscapes [20] is a widely used benchmark for image-based semantic segmentation. Since then, a number of large scale datasets for semantic segmentation have been released [21], [22], [23], [24]. The Waymo open dataset [25] and nuScenes [26] instead focus mainly on object detection or tracking. Oxford RobotCar [27] and KAIST Urban [28] target localization and mapping, while DDAD [29] and DrivingStereo [6] are specialized for monocular or stereo depth estimation respectively. KAIST MultiSpectral differs in that their datasets contains data from a thermal camera. Instead of capturing real-world data, Synthia [30] and [31] provide perfect ground truth in synthetic environments. However, all aforementioned datasets do not contain event camera data.

In this paragraph, we outline datasets in the automotive domain that contain event camera data. Table I provides an overview. The ADD dataset [32] contains a single low resolution (304x240) event camera for 2D object detection. This dataset was recently extended with a high resolution (1 MP) event camera [33] for the same task. N-CARS [34] contains very short (100 ms) sequences for binary classification of cars. DET [37] is a dataset captured with a CeleX-V DVS camera for driving lane extraction. Brisbane-Event-

Dataset	Frame Camera			Event Camera		Lidar	Groundtruth
	Resolution [MP]	Baseline [cm]	Color	Resolution [MP]	Baseline [cm]		
ADD [32]				0.1	-	-	2D Bounding Box
1MP Detection [33]	4	-	yes	0.9	-	-	2D Bounding Box
N-CARS [34]				0.01	-	-	Binary Class
DDD17 [35]	0.1	-	no	0.1	-	-	Vehicle Control, GPS
DDD20 [36]	0.1	-	no	0.1	-	-	Vehicle Control, GPS
DET [37]				1.0	-	-	Lane Extraction
Brisbane-Event-VPR [38]	2.1	-	yes	0.1	-	-	GPS, Place Matches
MVSEC [2]	0.4	10	no	0.1	10	VLP-16	Depth, GPS
<b>DSEC (Ours)</b>	1.6	<b>51</b>	yes	0.3	<b>60</b>	VLP-16	Depth, RTK GPS

TABLE I: Comparison of different event camera datasets in driving scenarios. Empty or missing entries indicate *not available* or *not applicable* respectively. ADD, N-CARS and DET do not feature a frame camera. MVSEC and this work are the only dataset with a stereo camera setup. Unlike other event camera datasets, our dataset provides a sufficient baseline and camera resolution for stereo matching in driving scenarios. Note that a baseline of above 50 centimeters is consistent with the KITTI datasets [3], [5] and DrivingStereo [6]

VPR [38] presents a driving dataset with a color DAVIS346 camera for place recognition. DDD17 [35] and DDD20 [36] are datasets comprising many hours of driving data from a monochrome DAVIS346 camera featuring various vehicle control data. These event camera datasets do not contain stereo event cameras and can thus not be used for the development of stereo matching algorithms.

Closest to our work is the MVSEC dataset [2]. It contains two DAVIS346B cameras with a resolution of roughly 0.1 MP and a baseline of 10 cm. Furthermore, the dataset features a stereo monochrome frame-based camera setup with a similar baseline and a resolution of 0.4 MP. What our dataset has in common with MVSEC is the LIDAR sensor. Both utilize a Velodyne VLP16 for the computation of ground truth depth data in post-processing. Different from MVSEC, however, we provide event camera data with 3 times higher resolution from two Prophesee Gen 3.1 sensors with a baseline of 60 cm similar to KITTI 360 [5]. In addition to that, we mount two RGB global shutter cameras each next to the event cameras with a baseline of 51 cm. The combination of significantly higher camera resolution and larger baseline therefore addresses one of the main limitations of the driving sequences of the MVSEC dataset.

### III. DATASET

#### A. Sensors

Our sensor suite consists of a multi-camera setup equipped with a Velodyne lidar and an RTK GPS system. We refer to figure 3 for visualization and table II for details of the characteristics of each sensor.

The monochrome stereo event cameras have a resolution of  $640 \times 480$  and are mounted with a baseline of 60 centimeters. Next to each event camera, a color FLIR Blackfly S USB3, with a resolution of  $1440 \times 1080$ , is mounted in a forward-facing direction. The baseline of 51 centimeters of the frame cameras is slightly smaller than the event camera baseline to account for the higher resolution of the Blackfly camera. We deliberately minimized the baseline between the two camera

Sensor Type	Description
2× Prophesee Gen3.1	Prophesee PPS3MVCD event camera 640 × 480 pixels, 3/4" CMOS Monochrome ≥ 120 dB dynamic range
2× FLIR Blackfly S USB3	FLIR BFS-U3-16S2C-CS 1440 × 1080 pixels, 1/2.9" CMOS Color 71.4 dB dynamic range 20 Hz frame rate
Velodyne VLP-16	16 Channels 100m range ±3cm range accuracy 10 Hz rotation rate
U-Blox RTK GNSS Receiver	ZED-F9P GNSS Receiver RTK signal with 10 Hz update rate ±1 cm position accuracy

TABLE II: Details about sensors utilized for data collection

types on each side of the stereo setup. By doing so, we achieved a baseline of approximately 4.5 centimeters on each side, which introduces only a small parallax and can be exploited for sensor fusion approaches.

The Blackfly cameras operate in auto exposure mode while prioritizing exposure time adjustments over increasing the analog gain to reduce noise. The maximum exposure time of 15 ms is typically only reached at night time, in which case the gain increases to minimize motion blur. The frames are captured in 12-bit raw format at 20 Hz. We apply a high-quality demosaicing algorithm [39] before mapping the debayered image to 8 bits.

The parameter settings, usually called biases, for event cameras differ fundamentally from standard cameras. Event cameras do not have an exposure setting. Instead, the sensitivity of the camera can be adjusted to be more responsive to changes in contrast. In bright daylight, we manually set the sensitivity to a higher value which means that the contrast threshold is lower. As a consequence, more events are generated and more detail in the scene is revealed. This is especially important in

direct sunlight where lens flare reduces the contrast. On the other hand, we increase the contrast threshold in dawn/dusk and night conditions to reduce the number of noisy events.

### B. Sequences

Split	Area	Time	#Sequences	Duration [s]
<b>Train</b>	Interlaken	Day	5	367
	Thun	Day	1	12
	Zurich City	Day, Night	35	2255
			<b>41</b>	<b>2634</b>
<b>Test</b>	Interlaken	Day	3	251
	Thun	Day	2	64
	Zurich City	Day, Night	7	244
			<b>12</b>	<b>559</b>
<b>Total</b>			<b>53</b>	<b>3193</b>

TABLE III: Details about the dataset sequences. In total, the dataset comprises close to an hour of driving data in diverse illumination conditions while driving in urban and rural environments.

Figure 8 illustrates the trajectories of the data recording. Over 4 TB raw data was recorded in urban, suburban, and rural areas in Switzerland. From the raw recordings we carefully select sequences with good recording and calibration quality. As summarized in table III, the final dataset consists of 53 sequences in three different areas of Switzerland at day and night time.

## IV. SYNCHRONIZATION AND CALIBRATION

This section describes the methodology for time synchronization of the sensors as well as the extrinsic and intrinsic calibration procedures. Time synchronization is achieved on hardware by using a microcontroller that associates GPS trigger and time signals with trigger signals sent to the cameras. The camera intrinsic and extrinsic calibration procedure uses E2VID, a neural network-based image reconstruction [40], in combination with Kalibr [41]. Finally, camera to lidar calibration is performed by manual rotational adjustment based on an accurate CAD model.

### A. Time Synchronization

For an overview of the time synchronization procedure, see Fig. 4. Time synchronization of the sensor suite was achieved through the use of hardware triggers supplied by the microcontroller (Fig. 4 MC) and GPS receiver. The GPS receiver first sends triggers in the form of PPS (pulse per second) signals at the top of each second to the microcontroller and the lidar, which update their internal timestamp accordingly. This means that lidar measurements are already provided in GPS timestamps. Upon receiving a PPS signal, the microcontroller forwards a sequence of 20 pulses to the events cameras and standard cameras. Each time the event camera receives a pulse it records a so-called “special event” which is precisely

timestamped with the clock of the event camera. When a pulse arrives at the standard camera, it triggers the camera to start exposing. Thus the standard camera is triggered at 20 Hz, and the event cameras receive a stream of special events at 20 Hz. For synchronization, we use the timestamps provided by the GPS triggers to compute the pulse times as they are transmitted by the microcontroller. Finally, we use the timestamps of the pulses to express frame timestamps, and event timestamps in the GPS clock.

### B. Camera Calibration

For the estimation of camera intrinsics and extrinsics, we use a checkerboard pattern together with the Kalibr toolbox [41]. While this procedure is straightforward with standard cameras, it is not directly applicable to event cameras. Instead of relying on blinking patterns, we approach the problem with image reconstruction [40] from the event streams. Figure 5 shows an example of reconstructed images from events that are directly used by Kalibr to detect corners and extract the checkerboard pattern. For accurate extrinsic calibration, it is important that all frames must be captured or reconstructed simultaneously. Hence, we can directly make use of our hardware time synchronization scheme to reconstruct images at the middle of the exposure times of the frames.

### C. Camera to Lidar Calibration

The CAD model of the recording setup provides an initial estimate of the transformations between the different sensors. To further refine the initial estimate, we employ refinement steps using ICP. We select  $N$  different views in a sequence of interest. For each pose,

- compute a local pointcloud from the Lidar measurements,
- compute the 3D pointcloud resulting from running the SGM [42] algorithm on the rectified image pair,
- align the stereo and Lidar pointcloud with a modified version of point-to-plane ICP [43] that only optimizes for rotation. We do not require ICP to optimize for translation because the translation can be accurately retrieved from the CAD model.

Finally, we average the  $N$  rotation quaternions [44] for a more robust extrinsic calibration.  $N$  is typically chosen between 2 and 5 but more poses can be used with this approach.

## V. GROUND TRUTH DEPTH GENERATION

The previously obtained intrinsic and extrinsic calibration can now be used to generate groundtruth depth maps. A single lidar scan only provides very sparse depth information. To improve the ground truth density, LIO-Mapping [45] computes the pose for each scan of the lidar. We query each pose and extract the relative poses of several preceding and proceeding scans. An example of such a pointcloud is shown in figure 6. Simply projecting this pointcloud into the camera frames is problematic because potential occlusions and moving objects are not taken into account. To address this issue we implement a filtering method similar to those employed by the KITTI depth [46] and DrivingStereo dataset [6]:

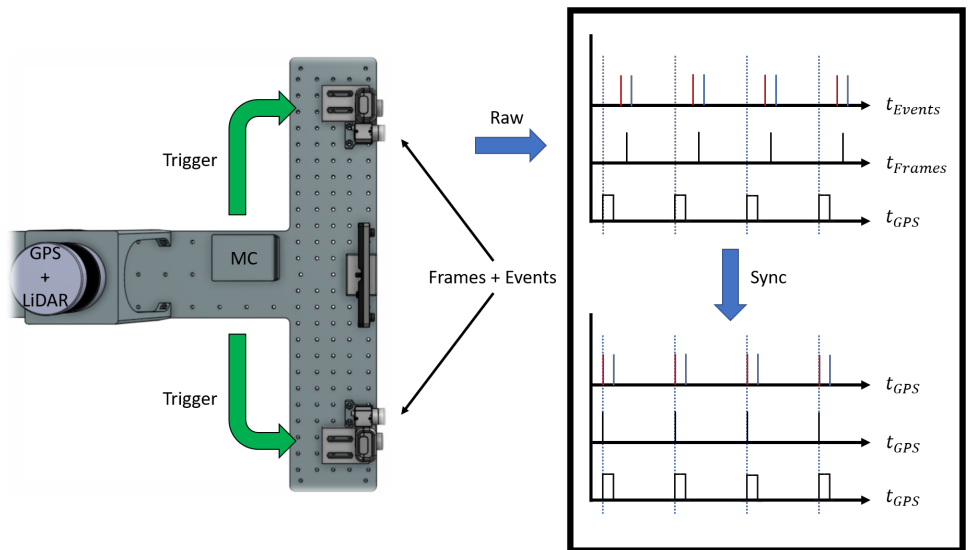


Fig. 4: The sensor suite used to record DSEC. It includes a GPS receiver (GPS), a Velodyne VLP16 (LIDAR), a microcontroller (MC), two Prophesee Gen3.1M Event Cameras (VGA resolution), and two FLIR BlackFly S USB3 RGB cameras. The microcontroller is synchronized to the GPS module and sends a trigger signal at 20 Hz to both event cameras and both standard cameras. The event cameras record these pulses in the form of special events, which are precisely timestamped and thus allow synchronization with the microcontroller clock. The standard camera starts the exposure time whenever a trigger signal arrives.

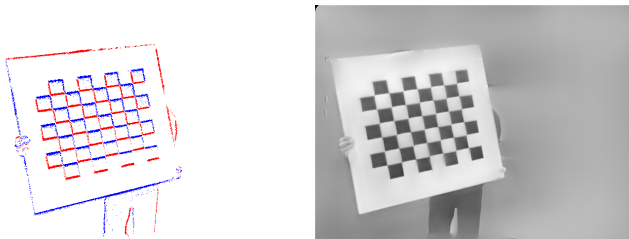


Fig. 5: Events (left) are used to reconstruct images (right) at mid-exposure times of the frames. These reconstructions can then be used in conjunction with standard calibration frameworks.

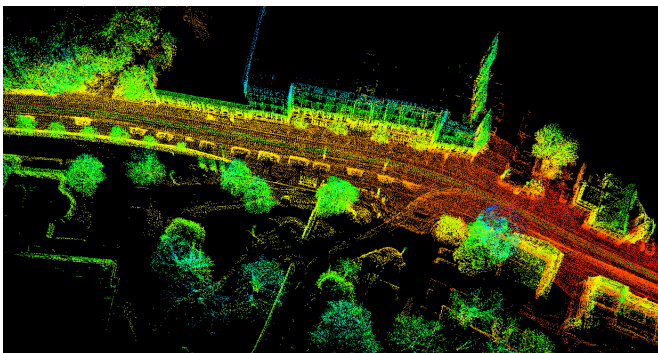


Fig. 6: A local pointcloud that is used to generate depth ground truth via projection into camera views.

- Accumulate a local Lidar pointcloud for each view
- Generate a disparity map from the rectified image pair using SGM [42] which is not affected by moving objects.
- The lidar pointcloud is projected into the camera frame.

Only the 3D points are retained that correspond to a disparity value close enough to the disparity at that pixel estimated by SGM. We remove 3D points further away than 50 meters to maintain an accurately filtered map.

- The filtered Lidar pointcloud is subsequently projected into the left event and frame cameras to generate the groundtruth disparity maps at 10 Hz.

Figure 7 visualizes the resulting groundtruth disparity maps without artefacts from moving objects or occluded points. As an example of filtered groundtruth due to moving objects is the top row in figure 7. Even though the front car is approximately 40 meters from the camera, the filtering approach successfully removes erroneous points.

## VI. EVALUATION METRICS

We evaluate disparity estimation methods using a number of established metrics in stereo matching. Following [3], [47], we evaluate methods on this dataset with the  $N$  pixel disparity error metric ( $DN$ ). This metric computes the percentage of disparity predictions with error larger than  $N$  pixels. We choose  $N$  as 2 and 1 to assess the predictions in both coarse and fine-grained manner. Furthermore, as in the Middlebury dataset, we also evaluate the mean absolute disparity error (MAE) and root-mean-square disparity error (RMSE) both in pixels.

## VII. EXPERIMENTS

In order to assess the difficulty of this dataset, the stereo method presented in [48] is trained on MVSEC and evaluated on the test set of DSEC. The resulting D1 error of 85.9 percent and MAE of 13.0 suggest that the model does not generalize to this dataset. Consequently, we retrain their model on the



Fig. 7: Examples of recorded event and image data with corresponding disparity groundtruth. From left to right: events with the corresponding disparity and image with corresponding disparity. The disparity image is colored to provide a sense of depth.

Area	Time	D1	D2	MAE	RMSE
<i>Interlaken</i>	Day	10.57	3.09	0.568	1.357
<i>Thun</i>	Day	10.84	3.13	0.621	1.595
<i>Zurich City</i>	Day	10.50	2.35	0.550	1.289
	Night	15.47	4.26	0.671	1.649
	Total	11.28	2.65	0.568	1.345
<b>Total</b>		<b>10.91</b>	<b>2.90</b>	<b>0.574</b>	<b>1.379</b>

TABLE IV: Baseline evaluation on the test set. We evaluate the method presented in [48] on a total of 12 sequences in 3 different areas within Switzerland. Here we report the average scores separated by city and day or night time as well as the total score. DN refers to the percentage of disparity errors higher than  $N$  pixels. MAE is mean absolute error and RMSE is the root-mean-square error.

training set for 200k iterations without data augmentation and choose the best performing for the final evaluation. Note that our baseline implementation uses the voxel grid representation [49] that contains events within a duration of 50 milliseconds. We empirically found that the performance is comparable to using event queues while significantly faster to train.

Table IV provides a summary of the results of the baseline experiment. We compute the errors for each area individually and also distinguish day and night sequences in the Zurich area. To assess the overall performance, we compute the mean of the error of all predictions. Evidently, the error metrics are comparable between the different locations. However, the D1 error is approximately 50% higher during night compared to the day sequences. This is to be expected because the signal to noise ratio decreases in low illumination conditions. Overall, the evaluation suggests that stereo methods trained on MVSEC do not generalize to this dataset and must be retrained to reach acceptable performance. Furthermore, we implemented and evaluated a baseline method on which future work can improve.

## VIII. DISCUSSION

During the recording sessions, we made several observations that might be of interest to users of this dataset.

Some of the provided sequences contain situations where the cameras face the sun directly, as shown in figure 3 or 3. In such scenarios, the contrast of the cameras is significantly reduced due to lens flare. We tuned the biases and settings of the event cameras to the best of our knowledge. Still, we believe that higher quality lenses and better lens coating could reduce the lens flare effect in the future.

On the other hand, we also provide sequences during night time. Night driving is challenging for standard and event cameras. Standard cameras, for example, rely on increasing the exposure time, which introduces a significant amount of motion blur. The event camera increasingly produces noisy events due to the reduced signal-to-noise ratio of the signal. An often overlooked difficulty for event cameras at night are flashing lights. Figure 9 highlights this effect and its influence on the spiking event rate while the cameras are in close proximity to a flashing street lamp. Rarely, the event readout of the event cameras could saturate and slightly delay the events.

## IX. LIMITATIONS

The groundtruth depth generation approach in section V, while handling occlusions and moving objects, introduces biases into the disparity groundtruth. First, the filtering approach leads to increased sparsity of disparity labels on moving objects. Second, the filtering approach is based on SGM and performed on the images. Hence, the resulting disparities are sparser in challenging conditions such as night driving or oversaturation of the images.

## X. CONCLUSION

This work proposes DSEC, the first large-scale event camera dataset in driving scenarios with a wide-baseline and high resolution event cameras. We additionally record RGB frames that are time-synchronized to the event cameras to foster

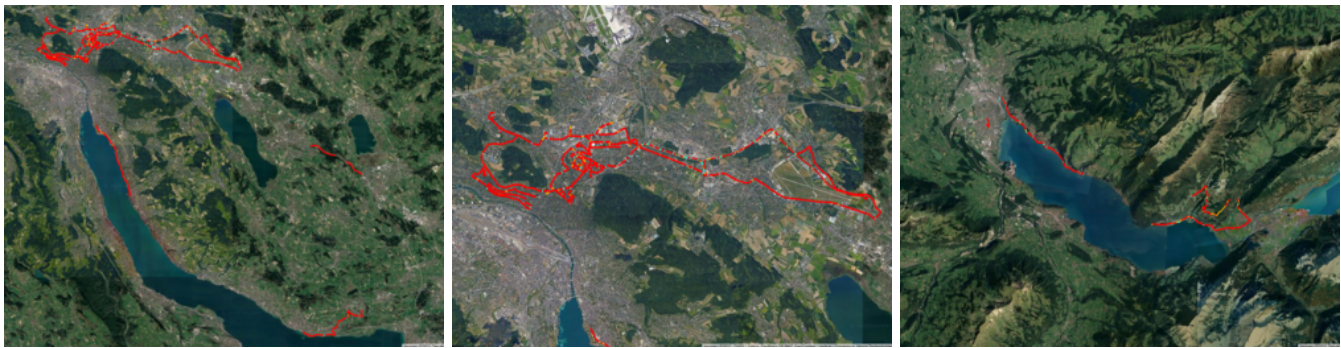


Fig. 8: Overlay of GPS trajectories onto a satellite map: The trajectories are colored according to the quality of the GPS reception. Red, orange and cyan indicates highest, medium or lower reception quality respectively. Data was recorded in various urban, suburban and rural areas in Switzerland. The left image shows the region around Lake Zurich. The center image is the section showing the trajectories around the city of Zurich while the right image shows the region of Lake Thun.

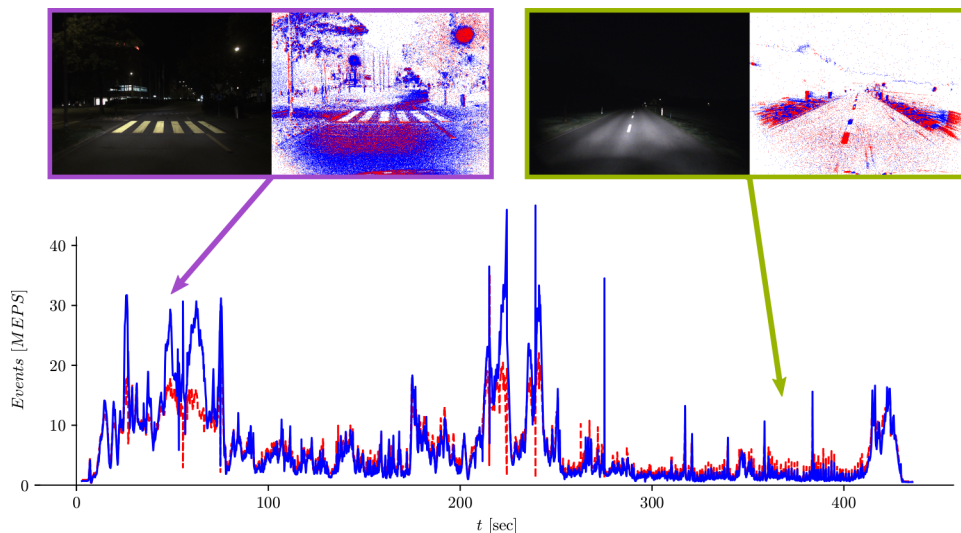


Fig. 9: The event rate in millions per second (MEPS) during a night sequence. The red and blue curves correspond to negative and positive event polarity respectively. The upper left pair of RGB and event frame corresponds to a situation where a street lamp flashes and periodically triggers events for a large number of pixels. In contrast, the upper right pair of RGB and event frame depicts a scene without artificial light sources other than the car. In this scenario, the event rate is much lower due to the absence of periodically flashing light sources.

research in sensor fusion. We believe that this dataset serves as a foundation to further explore the utility of event cameras in safety-critical applications such as autonomous driving. In the future, we plan to continuously extend the dataset with new tasks, improved hardware, and corner cases.

#### ACKNOWLEDGMENT

The authors would like to thank Mario Millhäusler for his contributions to the hardware design.

#### REFERENCES

- [1] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, “Event-based vision: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.
- [2] A. Z. Zhu, D. Thakur, T. Ozaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, “The multivehicle stereo event camera dataset: An event camera dataset for 3D perception,” *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2032–2039, July 2018.
- [3] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the KITTI vision benchmark suite,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2012.
- [4] M. Menze and A. Geiger, “Object scene flow for autonomous vehicles,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2015.
- [5] J. Xie, M. Kiefel, M.-T. Sun, and A. Geiger, “Semantic instance annotation of street scenes by 3d to 2d label transfer,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3688–3697.
- [6] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou, “Driving-stereo: A large-scale dataset for stereo matching in autonomous driving scenarios,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019, pp. 899–908.
- [7] D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys, “Variable baseline/resolution stereo,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. IEEE, 2008, pp. 1–8.
- [8] Y. Zhou, G. Gallego, H. Rebecq, L. Kneip, H. Li, and D. Scaramuzza, “Semi-dense 3d reconstruction with a stereo event camera,” in *Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 235–251.
- [9] S. Schraml, A. Nabil Belbachir, and H. Bischof, “Event-driven stereo matching for real-time 3d panoramic vision,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2015, pp. 466–474.
- [10] A. Z. Zhu, Y. Chen, and K. Daniilidis, “Realtime time synchronized

- event-based stereo,” in *Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 438–452.
- [11] A. Andreopoulos, H. J. Kashyap, T. K. Nayak, A. Amir, and M. D. Flickner, “A low power, high throughput, fully event-based stereo system,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018, pp. 7532–7542.
- [12] L. Steffen, D. Reichard, J. Weinland, J. Kaiser, A. Roennau, and R. Dillmann, “Neuromorphic stereo vision: A survey of bio-inspired sensors and algorithms,” *Frontiers in neurorobotics*, vol. 13, p. 28, 2019.
- [13] D. Gehrig, H. Rebecq, G. Gallego, and D. Scaramuzza, “EKLt: Asynchronous photometric feature tracking using events and frames,” *Int. J. Comput. Vis.*, 2019.
- [14] Z. Jiang, Y. Zhang, D. Zou, J. Ren, J. Lv, and Y. Liu, “Learning event-based motion deblurring,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 3320–3329.
- [15] C. Scheerlinck, N. Barnes, and R. Mahony, “Asynchronous spatial image convolutions for event cameras,” *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 816–822, Apr. 2019.
- [16] L. Pan, C. Scheerlinck, X. Yu, R. Hartley, M. Liu, and Y. Dai, “Bringing a blurry frame alive at high frame-rate with an event camera,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019.
- [17] L. Pan, M. Liu, and R. Hartley, “Single image optical flow estimation with an event camera,” *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020.
- [18] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, “A 240x180 130dB 3 $\mu$ s latency global shutter spatiotemporal vision sensor,” *IEEE J. Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, 2014.
- [19] T. Finatou, A. Niwa, D. Matolin, K. Tsuchimoto, A. Mascheroni, E. Reynaud, P. Mostafalu, F. Brady, L. Chotard, F. LeGoff, et al., “5.10 a 1280 $\times$  720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86  $\mu$ m pixels, 1.066 geps readout, programmable event-rate controller and compressive data-formatting pipeline,” in *2020 IEEE International Solid-State Circuits Conference-ISSCC*. IEEE, 2020, pp. 112–114.
- [20] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016.
- [21] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kotschieder, “The mapillary vistas dataset for semantic understanding of street scenes,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 4990–4999.
- [22] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, “The apollo-scape dataset for autonomous driving,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018, pp. 954–960.
- [23] J. Geyer, Y. Kassahun, M. Mahmudi, X. Ricou, R. Durgesh, A. S. Chung, L. Hauswald, V. H. Pham, M. Mühlegg, S. Dorn, et al., “A2d2: Audi autonomous driving dataset,” *arXiv e-prints*, 2020.
- [24] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, “Bdd100k: A diverse driving dataset for heterogeneous multitask learning,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 2636–2645.
- [25] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Cai, B. Caine, et al., “Scalability in perception for autonomous driving: Waymo open dataset,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 2446–2454.
- [26] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuscenes: A multimodal dataset for autonomous driving,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 11 621–11 631.
- [27] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 Year, 1000km: The Oxford RobotCar Dataset,” *Int. J. Robot. Research*, vol. 36, no. 1, pp. 3–15, 2017.
- [28] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, “Complex urban dataset with multi-level sensors from highly diverse urban environments,” *Int. J. Robot. Research*, vol. 38, no. 6, pp. 642–657, 2019.
- [29] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, “3d packing for self-supervised monocular depth estimation,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 2485–2494.
- [30] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, “The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016, pp. 3234–3243.
- [31] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, “Virtual worlds as proxy for multi-object tracking analysis,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016, pp. 4340–4349.
- [32] P. de Tournemire, D. Nitti, E. Perot, D. Migliore, and A. Sironi, “A large scale event-based detection dataset for automotive,” *arXiv e-prints*, pp. arXiv–2001, 2020.
- [33] E. Perot, P. de Tournemire, D. Nitti, J. Masci, and A. Sironi, “Learning to detect objects with a 1 megapixel event camera,” *arXiv e-prints*, 2020.
- [34] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman, “HATS: Histograms of averaged time surfaces for robust event-based object classification,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018, pp. 1731–1740.
- [35] J. Binas, D. Neil, S.-C. Liu, and T. Delbruck, “Ddd17: End-to-end davis driving dataset,” *arXiv e-prints*.
- [36] Y. Hu, J. Binas, D. Neil, S.-C. Liu, and T. Delbruck, “Ddd20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction,” *arXiv e-prints*, 2020.
- [37] W. Cheng, H. Luo, W. Yang, L. Yu, S. Chen, and W. Li, “Det: A high-resolution dvs dataset for lane extraction,” in *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, 2019, pp. 0–0.
- [38] T. Fischer and M. J. Milford, “Event-based visual place recognition with ensembles of temporal windows,” *IEEE Robotics and Automation Letters*, 2020.
- [39] D. Menon, S. Andriani, and G. Calvagno, “Demosaiicing with directional filtering and a posteriori decision,” *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 132–141, 2006.
- [40] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, “High speed and high dynamic range video with an event camera,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.
- [41] P. Furgale, J. Rehder, and R. Siegwart, “Unified temporal and spatial calibration for multi-sensor systems,” in *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2013.
- [42] H. Hirschmüller, “Stereo processing by semiglobal matching and mutual information,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2007.
- [43] Y. Chen and G. Medioni, “Object modelling by registration of multiple range images,” *Image and vision computing*, vol. 10, no. 3, pp. 145–155, 1992.
- [44] F. L. Markley, Y. Cheng, J. L. Crassidis, and Y. Oshman, “Averaging quaternions,” *Journal of Guidance, Control, and Dynamics*, vol. 30, no. 4, pp. 1193–1197, 2007.
- [45] H. Ye, Y. Chen, and M. Liu, “Tightly coupled 3d lidar inertial odometry and mapping,” in *IEEE Int. Conf. Robot. Autom. (ICRA)*. IEEE, 2019.
- [46] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, “Sparsity invariant cnns,” in *2017 international conference on 3D Vision (3DV)*. IEEE, 2017, pp. 11–20.
- [47] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, “High-resolution stereo datasets with subpixel-accurate ground truth,” in *German conference on pattern recognition*. Springer, 2014, pp. 31–42.
- [48] S. Tulyakov, F. Fleuret, M. Kiefel, P. Gehler, and M. Hirsch, “Learning an event sequence embedding for dense event-based deep stereo,” in *Int. Conf. Comput. Vis. (ICCV)*, October 2019.
- [49] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, “Unsupervised event-based learning of optical flow, depth, and egomotion,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019.