

# TUM-VIE: The TUM Stereo Visual-Inertial Event Dataset

Simon Klenk\*, Jason Chui\*, Nikolaus Demmel, Daniel Cremers

*Abstract*—Event cameras are bio-inspired vision sensors which measure per pixel brightness changes. They offer numerous benefits over traditional, frame-based cameras, including low latency, high dynamic range, high temporal resolution and low power consumption. Thus, these sensors are suited for robotics and virtual reality applications. To foster the development of 3D perception and navigation algorithms with event cameras, we present the TUM-VIE dataset. It consists of a large variety of handheld and head-mounted sequences in indoor and outdoor environments, including rapid motion during sports and high dynamic range scenarios. The dataset contains stereo event data, stereo grayscale frames at 20Hz as well as IMU data at 200Hz. Timestamps between all sensors are synchronized in hardware. The event cameras contain a large sensor of 1280x720 pixels, which is significantly larger than the sensors used in existing stereo event datasets. We provide ground truth poses from a motion capture system at 120Hz during the beginning and end of each sequence, which can be used for trajectory evaluation. TUM-VIE includes challenging sequences where state-of-the-art visual SLAM algorithms either fail or result in large drift. Hence, our dataset can help to push the boundary of future research on event-based visual-inertial perception algorithms.

## I. INTRODUCTION

Event cameras, also known as dynamic vision sensors (DVS), are passive imaging sensors, which report changes in the observed logarithmic brightness independently per pixel. Their main benefits are very high dynamic range (up to 140dB compared to 60dB of traditional cameras), high temporal resolution and low latency (in the order of microseconds), low power consumption and strongly reduced motion blur [1, 2]. Hence, these novel sensors have the potential to revolutionize robotic perception. Figure 1 shows the superior performance of an event camera in low light conditions.

Due to the novelty of the field, algorithms for event-based sensors are still immature compared to frame-based algorithms. Computer vision research in the last fifty years has strongly profited from publicly available datasets and benchmarks. To advance research with these novel and expensive sensors, we introduce the TUM-VIE dataset. It contains a variety of sequences captured by a calibrated and hardware-synchronized stereo pair of Prophesee GEN4 CD sensors with 1280x720 pixels resolution. To our knowledge, TUM-VIE is the first stereo dataset featuring such a high-resolution event camera, surpassing other datasets by at least a factor of ten in the number of event pixels per camera. In principle, this allows for more detailed reconstructions.

\* These authors contributed equally. The authors are with the Computer Vision Group, Department of Informatics, Technical University of Munich, Germany. {simon.klenk, jason.chui, nikolaus.demmel, cremers}@tum.de.

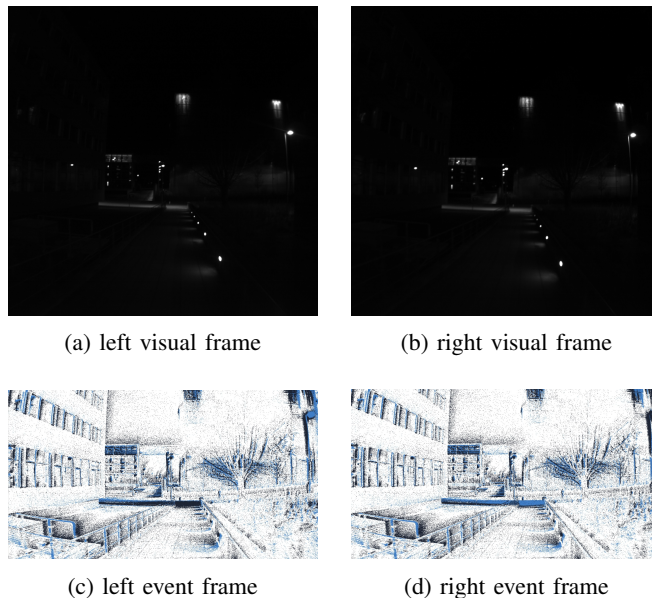


Fig. 1: This scene is captured in low-light conditions outside. It demonstrates the high dynamic range advantage of the event camera (bottom row) compared to the visual camera (top row). It shows all four camera modalities of the sequence **bike-dark** at 83.6 seconds recording time. Note that later in this sequence there is more ambient light present and hence the visual frames can still be used in computer vision algorithms. For this visualization, the asynchronous events are accumulated into frames using an integration time of 5 milliseconds.

More importantly, it enables the evaluation of frame-based versus event-based algorithms on data with similar, state-of-the-art resolution. To facilitate this, our setup contains two hardware-synchronized global shutter cameras of resolution 1024x1024 pixels, 12 bit color depth, known exposure times and vignette calibration. The IMU provides 3-axis accelerometer and gyroscope data at 200 Hz. Contrary to most existing datasets, we provide the calibration of IMU biases, as well as axis scaling and misalignment similar to [3]. The timestamps of all sensors are synchronized in hardware.

TUM-VIE can for example be used to develop algorithms for localization and mapping (visual odometry and SLAM), feature detection and tracking, 3D reconstruction, as well as self-supervised learning and sensor fusion. The availability of comprehensive datasets combining comparable sensor modalities to TUM-VIE is quite small. Furthermore, sensor

characteristics of event camera have greatly improved within the last decade [1]. Hence, we believe it is important to use the most recent event camera for a meaningful comparison with frame-based algorithms.

The main contributions of this paper are:

- We present the first 1 megapixel stereo event dataset featuring IMU data at 200Hz and stereo global shutter grayscale frames at 20Hz with known photometric calibration and known exposure times. Timestamps between all sensors are hardware-synchronized.
- The first stereo event dataset featuring head mounted sequences (relevant for VR) and sport activities with rapid and high-speed motions (biking, running, sliding, skateboarding).
- We propose a new method for event camera calibration using Time Surfaces.
- We evaluate our dataset with state-of-the-art visual odometry algorithms. Furthermore, we make all calibration sequences publicly available.

The dataset can be found at:

<https://go.vision.in.tum.de/tumvie>

## II. RELATED WORK

Weikersdorfer et al. [4] present one of the earlier event datasets, combining data from an eDVS of 128x128 pixels and an RGB-D sensor in a small number of indoor sequences. They provide ground truth poses from a motion capture (MoCap) system, but the total sequence length only amounts to 14 minutes.

Barranco et al. [5] provide a dataset which focuses on evaluation of visual navigation tasks. They use a dynamic and active pixel vision sensor (DAVIS) which combines event detection alongside regular frame-based pixels in the same sensor. However, only 5 degree-of-freedom (DOF) motions are captured and the ground truth poses are acquired from wheel odometry which is subject to drift.

The work by Mueggler et al. [6] captures full 6 DOF motions and precise ground truth by a MoCap system indoors. The sequences include artificial scenes such as geometric shapes, posters or cart boxes, but also an urban environment and an office. Hardware-synchronized grayscale frames and IMU measurements are provided in addition to the events. They also present an event camera simulator in their work. However, the monocular DAVIS 240C merely captures 240x180 pixels and the total sequence length only amounts to 20 minutes.

There also exists a small number of related datasets targeting automated driving applications. Binas et al. present DDD17 [7] as well as the follow-up dataset DDD20 [8] which feature monocular event data from a DAVIS 346B with 346x260 pixels. The camera is mounted on a car windshield while driving through various environments and conditions, for a total of 12 and 51 hours, respectively. Additionally, GPS data and vehicle diagnostic data such as steering angle and vehicle speed are provided. Similarly, the dataset by Perot et al. [9] comprises 14 hours of recording from a car. It includes labelled bounding boxes for the classes of car, pedestrian

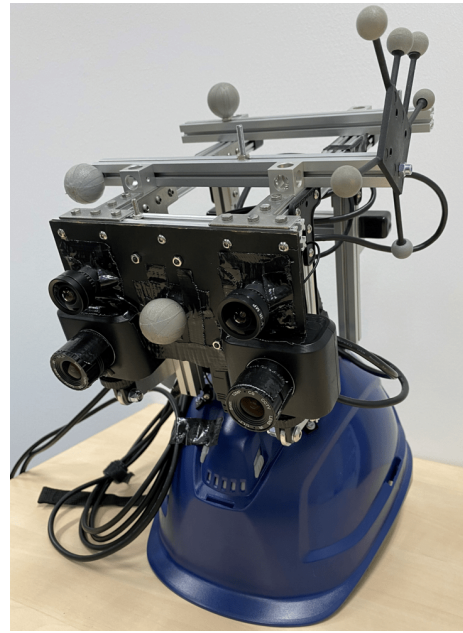


Fig. 2: Sensor setup mounted onto a helmet. The distance between event camera (bottom) and visual camera (top) is only 4.49 centimeters for the left and right side of the stereo rig. The baseline between the visual cameras is 10.86 centimeters, the baseline between the event cameras is 11.84 centimeters. These numbers are obtained from our calibration algorithm and confirmed by physical measurements.

and two-wheeler. They provide RGB images recorded at 4 megapixels as well as an event stream recorded by the 1 megapixel Prophesee GEN4-CD sensor, which is also used in this work.

Furthermore, there exist specialised datasets such as the UZH-FPV Drone Racing Dataset [10], which is targeted for localization of drones during high speed and high accelerations in 6 DOF. The sensor setup includes a miniDAVIS346 as well as a fisheye stereo camera with 640x480 pixels resolution and hardware-synchronized IMU. In addition, external ground truth poses of the drone are provided by a laser tracker system at 20Hz. However, partial tracking failures during high-acceleration maneuvers are reported. Lee et al. [11] present the dataset ViViD, which contains sequences for visual navigation in poor illumination conditions. In addition to lidar and RGB-D data, they record thermal images with a resolution of 640x512 pixels at 20Hz. They provide ground truth poses from a MoCap system indoors and use a state-of-the-art lidar SLAM system for outdoor ground truth poses. However, the monocular DAVIS stream features only a resolution of 240x180 pixels and the timestamps between the sensors are not synchronized in hardware.

The dataset MVSEC by Zhu et al. [12] as well as the recent dataset DSEC by Gehrig et al. [13] are most closely related to our work. Both MVSEC and DSEC contain a lidar sensor for depth estimation, two frame cameras, two event cameras as well as a GNSS receiver. DSEC addresses the limitation of a small camera baseline in MVSEC targeting automated

driving scenarios. TUM-VIE comes with a number of complementary advantages. First, our event camera provides 10 times more pixels than the DAVIS m346B in MVSEC and 3 times more pixels than the Prophesee Gen3.1 in DSEC. This can help to study the impact of high resolution data on algorithmic performance and in principle allows a more detailed reconstruction. Second, our visual stereo cameras encompass a largely increased field of view compared to the VI sensors in MVSEC. Third, in contrast to both MVSEC and DSEC, we provide a photometric calibration for our visual cameras, which is beneficial for employing direct VO methods on the intensity images [14]. Fourth, we provide calibrated IMU noise parameters which are required for accurate probabilistic modeling in state estimation algorithms [3]. The MVSEC dataset features sequences on a hexacopter, on a car at low speed (12 m/s), and on a motorcycle at high speed (12-36 m/s) totalling about 60 minutes. However, only about 6 minutes of handheld indoor recordings are provided. Similarly, the DSEC dataset contains 53 minutes of outdoor driving on urban and rural streets. Contrary to that, the TUM-VIE dataset provides extensive handheld and head-mounted sequences during walking, running and sports in indoor and outdoor scenarios and under various illumination conditions, totalling 48 minutes excluding the calibration sequences.

### III. DATASET

#### A. Sensor Setup

Table I gives an overview of our sensors and their characteristics. The full setup including the attached infrared markers can be seen in Figure 2. We use the Prophesee GEN4-CD Evaluation Kit which includes a 1 megapixel event sensor in a robust casing and a USB 3.0 interface. The Prophesee GEN4 CD sensor can handle a peak rate of 1 Giga-event per second. In most sequences we have a mean event rate of a few Mega-events per second.

The MoCap system emits 850nm infrared (IR) light in order to track the IR-reflective passive markers which are attached to the rigid body setup. Since the Prophesee GEN4 CD sensor is sensitive to IR light, moving the setup inside the MoCap room would create undesirable noise events. Therefore, we place an IR-blocking filter with cutoff frequency of 710nm parallel to the sensor, at a distance of approximately 2 millimeter, see Figure 3. The placement close to the sensor has the benefit that stray light inside the camera is also blocked.

#### B. Sequence Description

Table II gives an overview of the available sequences. Total recording length amounts to 48 minutes excluding the calibration sequences. For each day of recording, we provide a sequence called **calib** which is used to calibrate the extrinsic and intrinsic parameters of all cameras. The sequence called **imu-calib** is used to obtain the transformation between the marker coordinate system (infrared markers for MoCap) and the IMU, as well as to determine the IMU biases and scaling factors. Furthermore, we provide two

TABLE I: Hardware Specification

Sensor	Rate	Properties
2x Prophesee GEN4-CD lens: Foctek M12-5IR	$\leq 10^9 \frac{\text{events}}{\text{s}}$	1280x720 pixels FOV: 90°H / 65°V up to 124 dB
2x IDS Camera uEye UI-3241LE-M-GL lens: Lensagon BM4018S118	20 Hz	1024x1024 pixels FOV: 101°H / 76°V up to 60 dB global shutter
IMU Bosch BMI160 integrated on Genuino 101	200 Hz	3D accelerometer 3D gyroscope up to 60 dB temperature
MoCap OptiTrack Flex13	120 Hz	accurate 6D pose 850nm IR light



Fig. 3: The infrared filter is mounted parallel to the sensor, at a distance of approximately 2 millimeter. The sensor plane is located approximately at the 5 millimeter mark of the ruler.

sequences called **calib-vignette**, which are used to compute the photometric calibration of the visual cameras.

The first seven sequences in Table II with prefix **mocap** contain ground truth poses for the whole trajectory: **mocap-1d-trans** contains a simple one dimensional translational motion, **mocap-3d-trans** contains a three dimensional translation and **mocap-6dof** contains a full 6 DOF motion. All three sequences show a table containing diverse objects such as books, multiple similarly looking toy figures and cables. Most of the the scene is bounded by a calibration pattern placed behind the table. The sequences **mocap-desk** and **mocap-desk2** show a loop motion around two different office desks: **mocap-desk** shows two computer screens, a keyboard with some cables and the scene is bounded by a close-by white wall; **mocap-desk2** also shows two screens but the depth is less strictly bounded and there are also multiple calibration patterns and desk accessories visible.

The sequence **office-maze** is recorded during a walk through various offices and hallways in the university building. The sequence **running-easy** is recorded in handheld mode while running through the corridor of the office. In the sequence **running-hard**, the camera is rapidly rotated during running such that it faces the office wall for short moments. This makes it hard to perform camera tracking. However, the wall features research posters such that there is still texture present for tracking.

TABLE II: Sequence Overview

Sequence	Duration(s)	MER( $10^6$ events/s)
mocap-1d-trans	36.6	14.8
mocap-3d-trans	33.2	24.9
mocap-6dof	19.5	27.15
mocap-desk	37.5	29.7
mocap-desk2	21.4	28.4
mocap-shake	26.3	26.55
mocap-shake2	26.7	22.3
office-maze	160	28
running-easy	73	27.25
running-hard	72	26.2
skate-easy	79	26.25
skate-hard	86	25.7
loop-floor0	284	29.55
loop-floor1	257	29.55
loop-floor2	240	28.3
loop-floor3	256	29.7
floor2-dark	152	17.45
slide	196	28
bike-easy	288	26.8
bike-hard	281	26.9
bike-dark	261	25.35

The sequence **skate-easy** traverses the same corridor as running-easy and running-hard with a skateboard. In **skate-hard**, the camera is rapidly rotated to face the wall for a few seconds during the ride.

The sequence **loop-floor0** to **loop-floor3** are obtained by walking through the university building on the respective floor. These four datasets can be used to test loop closure detection algorithms in the event stream, which is still an immature research field. We additionally provide the sequence **floor2-dark**. This sequence shows the high dynamic range advantage of the event camera over the visual camera. We believe that with better event-based SLAM algorithms, the path could be tracked accurately even in this low-light condition, whereas state-of-the-art visual SLAM systems fail, see Table IV.

The sequence **slide** shows a path through the university building while sliding from floor3 to floor0 in the middle of the sequence. This sequence contains high dynamic range and high speed motion.

The bike sequences are recorded with the helmet worn on the head, whereas all other sequences are recorded in handheld mode. **bike-easy** contains a biking sequence during the day at medium speed. **bike-hard** traverses the same path as bike-easy, additionally containing rapid rotations of the sensor setup, suddenly facing sideways or down on the road. **bike-dark** contains a slightly shorter path than the other bike sequences, recorded at night in low-light conditions outside.

In Figure 1, 4 and 5 we show both visual frames and events of a few selected sequences. The events are visualized as accumulated frames with an accumulation time of 5 milliseconds. Positive events are visualized in blue, negative events in black and white color is used to indicate no change.

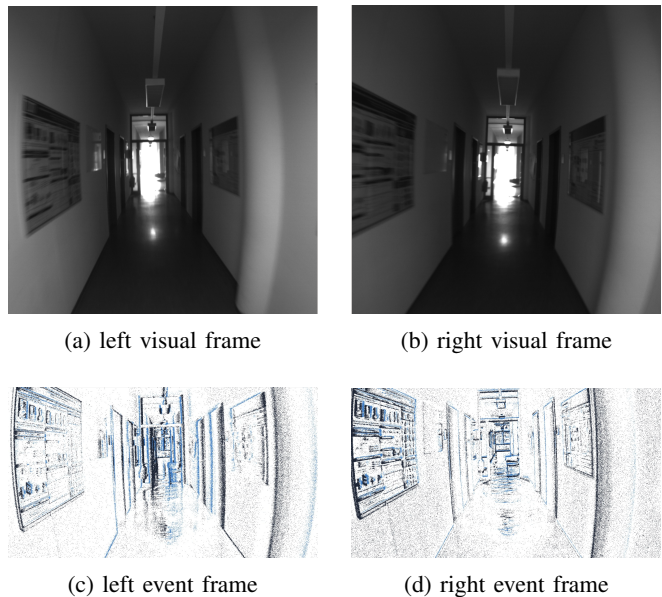


Fig. 4: **run-easy** sequence at 25.1 seconds. The visual frames contain significant motion blur on the poster and oversaturation in the middle part of the image, whereas the event frames can capture many details.

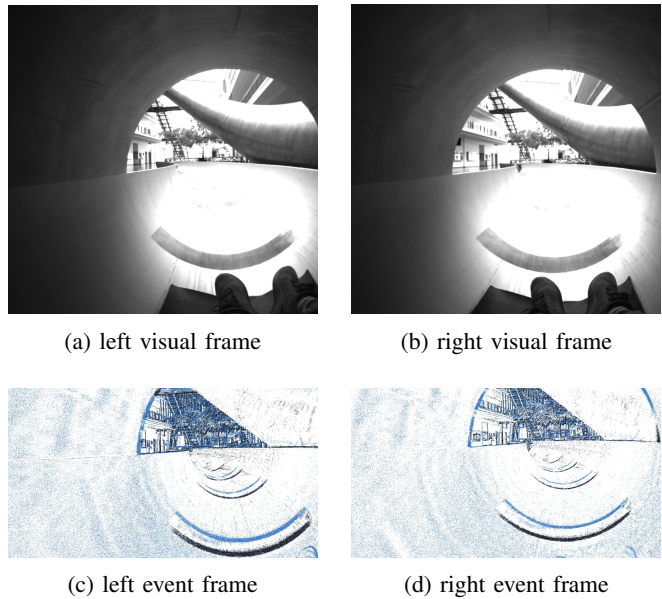


Fig. 5: **slide** sequence at 93.7 seconds. Due to the auto-exposure algorithm in the visual cameras, the exposure adapts quickly and details in the bright background are visible. However, a few frames earlier in this sequence, the frames are similarly oversaturated as in Figure 4.

## IV. CALIBRATION

### A. Intrinsic and Extrinsic Camera Calibration

The field of view among all cameras is only partially overlapping. Hence, to allow for extrinsic calibration between our camera modalities we use a pattern of AprilTags [15]. This pattern allows to perform robust data association between

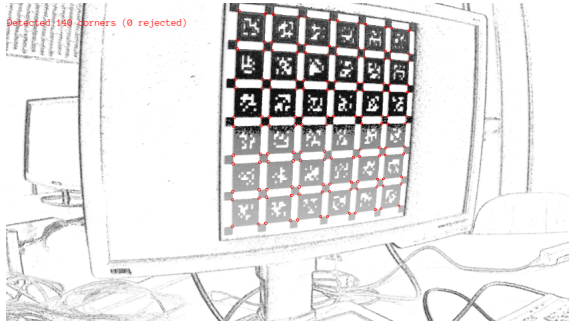


Fig. 6: The April-grid pattern has been reconstructed based on creating the time surface. Using time surfaces showed slightly better AprilTag detection recall compared to accumulated event frames.

recognized tags from different cameras. It is displayed by a commodity computer screen, blinking at a fixed frequency. We move the multi-camera setup around the screen to capture the pattern from various poses.

The intrinsic and extrinsic parameters for both visual and event cameras are determined in a joint optimization. To allow for accurate intrinsic parameter estimation, in particular radial distortion of the lenses, we validate for each calibration sequence that the detections of AprilTags are spread over the full image dimension of each camera.

To enable detection of AprilTags in the event stream, we transform the events into a frame-like structure. One popular method is to accumulate events within a certain time interval  $t \in [t_{target} - \Delta t, t_{target} + \Delta t]$  into frames. However, in order to create sharper event frames at time  $t_{target}$ , we use the time surface representation instead. Inspired by [16], we use time surfaces which take the latest as well as the next future event at time  $t_{target}$  into account. The time-surface  $\mathcal{T}_{t_{target}}$

is created as follow,

$$\mathcal{T}_{t_{target}}[x, y] = \frac{1}{2}(e^{-\alpha|t_{last}-t_{target}|} + e^{-\alpha|t_{next}-t_{target}|}) \quad (2)$$

where  $t_{last}$  is the timestamp of the previous event at pixel  $[x, y]$  before  $t_{target}$ ,  $t_{next}$  is the timestamp of next event at pixel  $[x, y]$  after  $t_{target}$  and  $\alpha$  is the decay rate. Using a time surface representation, we noticed a higher detection recall of AprilTags and hence a slightly more robust calibration. To perform this joint calibration, we modified the calibration tools from Basalt [17]. A time surface created with the described method can be seen in Figure 6. We provide the calibration parameters as well as the raw calibration sequences for each day of recording.

### B. IMU intrinsics, IMU-Camera-extrinsics

Similar to the TUM-VI dataset [3], we assume that the IMU measurements are subject to white noise with standard deviation  $\sigma_w$  and an additive bias value. The bias value is changing according to a random walk over time. The random walk is modelled as integration of white noise with standard deviation  $\sigma_b$ . To determine the parameter  $\sigma_w$ , a slope of  $-\frac{1}{2}$  is fitted to the range of  $0.02 \leq \tau \leq 1$  in the log-log plot of the Allan deviation. To determine the bias parameter  $\sigma_b$ , a slope of  $\frac{1}{2}$  is fitted to the range  $1000 \leq \tau \leq 6000$  seconds. The values for  $\sigma_w$  and  $\sigma_b$  can then be taken as the y-value of the fitted line at  $\tau = 1$  and  $\tau = 3$ , respectively. This procedure is visualized in Figure 7 and 8.

The extrinsic calibration between IMU and visual cameras is obtained by using a static Aprilgrid pattern. Calibration data is found in **imu-calib**. All six degrees of freedom are excited during this calibration while the pattern is constantly in the field of view of the visual cameras. The exposure time is set to a low value of 3.2 milliseconds to minimize motion blur but still allow detection of AprilTags. We leave the extrinsic between the visual cameras fixed and only optimize the transformation between IMU and visual camera frame.

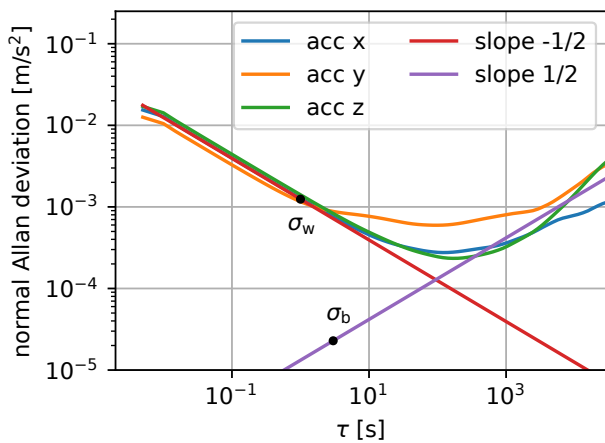


Fig. 7: Accelerometer log-log plot of Allan deviation over integration time  $\tau$ . We obtain the IMU noise values of  $\sigma_{w,acc} = 0.0012$  at  $\tau = 1$  and  $\sigma_{b,acc} = 2.2856 \times 10^{-5}$  at  $\tau = 3$ .

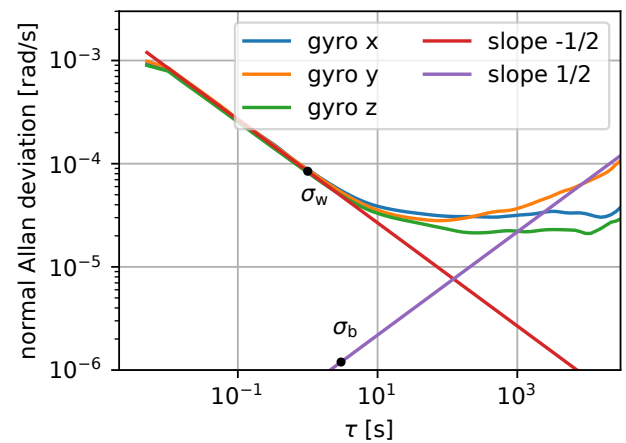


Fig. 8: Gyroscope log-log plot of Allan deviation over integration time  $\tau$ . We obtain the IMU noise values of  $\sigma_{w,gyro} = 8.4503 \times 10^{-5}$  at  $\tau = 1$  and  $\sigma_{b,gyro} = 1.1961 \times 10^{-6}$  at  $\tau = 3$ .

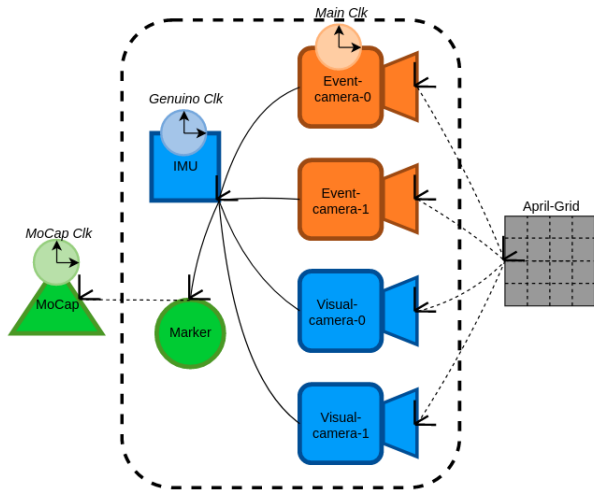


Fig. 9: Coordinate systems which we have used in this paper. Solid lines are the relative poses found during calibration, dashed lines are the relative poses measured in real-time. Colors indicate the respective raw sensor clock before temporal calibration.

### C. Temporal Calibration

All sensors on the setup are synchronized in hardware. An overview of the different sensor clocks can be seen in Figure 9. We report all timestamps in the left event camera’s clock (main clock). The synchronization between left and right event camera is achieved through the available synchronization connections of the Evaluation Kit.

The Genuino 101 triggers the visual cameras, measures exposure start and stop timestamps and reads the IMU values. As mentioned above, we define the timestamp of an image as mid-exposure point. Hence, timestamps of the images are well-aligned with IMU measurements in Genuino clock. However, due to the IMU readout delay, there exists a constant offset of 4.77 milliseconds. We determine this offset exactly once during calibration and correct it in all recordings.

To transform IMU and image timestamps from Genuino clock to main clock, we additionally measure exposure start and stop timestamps via trigger signal in the left event camera. Data association between triggers on the Genuino and triggers on the event camera is achieved through a distinctive startup sequence, which alternates extremely high and low exposure times. We notice a linear relation between Genuino and main clock, i.e. there exists a constant offset  $b$  and a small constant clock drift  $m$  between both clocks. Whereas the value of  $b$  is different for each sequence, the value of  $m$  is usually around 2 milliseconds per minute (Genuino clock runs faster than main clock). We obtain the linear coefficients from a least-squares fit and correct for them in each sequence, such that all sensor data is provided in main clock.

Ground truth poses are recorded in MoCap clock. We model a constant offset  $\Delta t_{gt,m}$  between MoCap clock and main clock. Since all data is recorded on the same com-

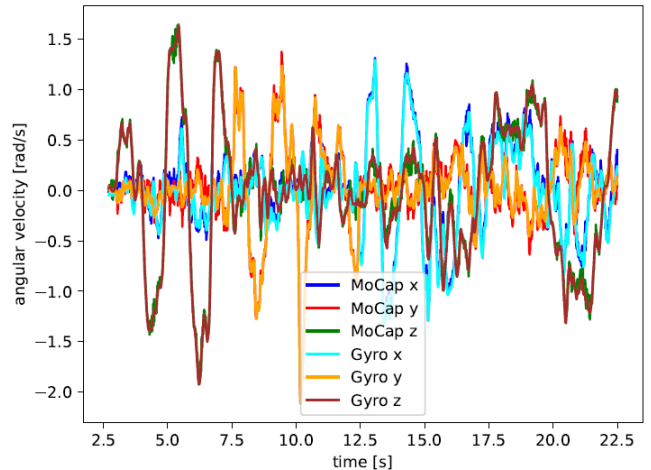


Fig. 10: Alignment of angular velocities between gyroscope measurements and Mocap estimations (using central differences of poses). The obtained time offset  $\Delta t_{gt,m}$  between MoCap clock and IMU clock is provided for each sequence.

puter, a rough estimate of this offset is obtained from the system’s recording time. A refinement of this estimate is obtained by aligning angular velocities measured by the IMU with estimated angular velocities which are obtained from central differences between MoCap poses, as can be seen in Figure 10. The offset  $\Delta t_{gt,m}$  which minimizes the sum of all squared errors is estimated for each sequence. For the trajectories where we exit and later re-enter the MoCap room, we compute this offset once for the beginning and once for the end of the sequence to account for clock drift. The provided ground truth poses in the dataset have corrected timestamps and are reported in the main clock. Additionally, we provide the estimated offsets to facilitate custom time-alignment approaches.

### D. Biases and Event Statistics

Table III shows the biases of the event cameras, which are the same in all sequences. The ratio of ON over OFF pixels is approximately 0.7 in all sequences and for both cameras.

## V. EVALUATION OF STEREO-VIO ALGORITHMS

To verify the data and calibration quality, we test all sequences with state-of-the-art open-source visual-inertial odometry systems. We provide evaluation for Basalt [17] and VINS-Fusion [18, 19, 20]. The methods are used with default

TABLE III: Biases of Prophesee GEN4-CD sensor

bias name	value
bias_diff	69
bias_diff_off	52
bias_diff_on	112
bias_fo.n	23
bias_hpf	48
bias_pr	151
bias_refr	45

TABLE IV: Absolute trajectory errors (ATE) [m]

Sequence	Basalt	VINS-Fusion	length[m]
mocap-1d-trans	0.003	0.011	5.01
mocap-3d-trans	0.009	0.010	6.85
mocap-6dof	0.014	0.017	5.30
mocap-desk	0.016	0.058	9.44
mocap-desk2	0.011	0.013	5.34
mocap-shake	x	x	24.5
mocap-shake2	x	x	26.2
office-maze	0.64	4.40	205
running-easy	1.34	0.78	113
running-hard	1.03	1.74	117
skate-easy	0.22	1.74	114
skate-hard	1.78	0.97	113
loop-floor0	0.58	3.43	358
loop-floor1	0.66	1.72	338
loop-floor2	0.48	2.04	282
loop-floor3	0.51	8.36	328
floor2-dark	4.54	4.54	254
slide	1.54	2.44	248
bike-easy	1.62	13.10	788
bike-hard	2.01	9.88	784
bike-dark	6.26	20.20	670

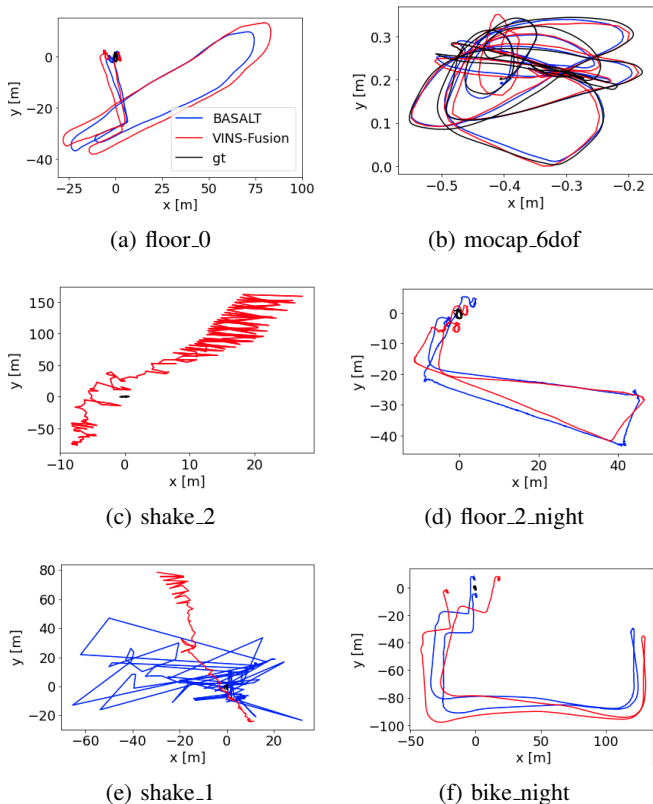


Fig. 11: Results of evaluated methods for 2 easy sequences (a, b) and 4 challenging sequences (c, d, e, f) from our dataset. The ground truth is shown in black for the segments of the trajectory where it is available. The presented results are obtained with default parameters. Noise parameters are set to inflated values from the Allan plots in Figure 7 and 8 to account for unmodeled noise and vibrations.

parameters on full resolution images (1024 x 1024 pixels). The results are summarised in Table IV and a visualization for some sequences is showed in Figure 11. An x in Table IV indicates an ATE larger than the sequence’s path length or that the method fails. Basalt and VINS-Fusion perform well for non-challenging sequences as expected.

However, our evaluation also shows that Basalt and VINS-Fusion result in large drift for most of the challenging sequences, e.g. with fast motion in **running-easy**, **running-hard**, **skate-hard**, **slide**, and **bike-hard** or with low light, e.g. in **floor2-dark** and **bike-dark**. This means that the dataset is challenging enough for state-of-the-art visual-inertial systems and can be used for further research in event-based visual-inertial odometry algorithms.

## VI. CONCLUSION

In this paper, we propose a novel dataset with a diverse set of sequences, including small and large-scale scenes. We specifically provide challenging sequences in low light and high dynamic-range conditions as well as during fast motion. Our dataset composes of a high-resolution event stream and images captured by a wide field of view lens, as well as IMU data and partial ground truth poses.

We evaluate our dataset with state-of-the-art visual-camera-based stereo-VIO. The results show that there are open challenges which need new algorithms and new sensors to tackle them. Event-based odometry algorithms are still immature compared to frame-based methods, which makes it difficult to present an evaluation of event-based algorithms. Hence, our dataset can be useful for further research in the development of event-based visual-inertial odometry, as well as 3D reconstruction and sensor fusion algorithms.

## ACKNOWLEDGMENT

We thank everybody who contributed to the paper. In particular, we want to thank Mathias Gehrig (University of Zurich) for supporting us with the h5 file format. We also thank Lukas Koestler (Technical University of Munich) for helpful suggestions and proofreading the paper. Additionally, we thank Yi Zhou (HKUST Robotics Institute) for his helpful comments. This work was supported by the ERC Advanced Grant SIMULACRON.

## REFERENCES

- [1] Guillermo Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, Stefan Leutenegger, A. Davison, J. Conradt, Kostas Daniilidis, and D. Scaramuzza. “Event-based Vision: A Survey”. In: *IEEE transactions on pattern analysis and machine intelligence* PP (2020).
- [2] Tobi Delbruck, Yuhuang Hu, and Zhe He. *V2E: From video frames to realistic DVS event camera streams*. 2020. arXiv: 2006.07722.
- [3] D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stueckler, and D. Cremers. “The TUM VI Benchmark for Evaluating Visual-Inertial Odometry”. In: *International Conference on Intelligent Robots and Systems (IROS)*. Oct. 2018.

- [4] D. Weikersdorfer, D. B. Adrian, D. Cremers, and J. Conradt. "Event-based 3D SLAM with a depth-augmented dynamic vision sensor". In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. 2014, pp. 359–364.
- [5] Francisco Barranco, Cornelia Fermuller, Yiannis Aloimonos, and Tobi Delbruck. "A Dataset for Visual Navigation with Neuromorphic Methods". In: *Frontiers in Neuroscience* 10 (2016), p. 49.
- [6] 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". In: *International Journal of Robotics Research*, Vol. 36, Issue 2, (Feb. 2017), pp. 142–149.
- [7] Binas Binas, Neil J., Liu D., S.-C., and T. Delbruck. "DDD17 End-To-End DAVIS Driving Dataset". In: *ICML17 Workshop on Machine Learning for Autonomous Vehicles*. 2017.
- [8] 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". In: *Special session Beyond Traditional Sensing for Intelligent Transportation (23rd IEEE International Conference on Intelligent Transportation Systems)*. Sept. 2020.
- [9] E. Perot, P. de Tournemire, D. Nitti, J. Masci, and A. Sironi. "Learning to Detect Objects with a 1 Megapixel Event Camera". In: *Advances in Neural Information Processing Systems 33 (NeurIPS)*. 2020.
- [10] Jeffrey Delmerico, Titus Cieslewski, Henri Rebecq, Matthias Faessler, and Davide Scaramuzza. "Are We Ready for Autonomous Drone Racing? The UZH-FPV Drone Racing Dataset". In: *IEEE Int. Conf. Robot. Autom. (ICRA)*. 2019.
- [11] A Lee, Cho J., Yoon Y., S. Shin, and A Y. Kim. "ViViD: Vision for Visibility Dataset". In: *IEEE Int. Conf. Robotics and Automation (ICRA) Workshop: Dataset Generation and Benchmarking of SLAM Algorithms for Robotics and VR/AR*. 2019.
- [12] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. "The Multi Vehicle Stereo Event Camera Dataset: An Event Camera Dataset for 3D Perception". In: *IEEE Robotics and Automation Letters* 3 (2018).
- [13] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. "DSEC: A Stereo Event Camera Dataset for Driving Scenarios". In: *IEEE Robotics and Automation Letters* (2021).
- [14] J. Engel, V. Koltun, and D. Cremers. "Direct Sparse Odometry". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Mar. 2018).
- [15] E. Olson. "Apriltag: A robust and flexible visual fiducial system," in: *2011 IEEE International Conference on Robotics and Automation*. 2011.
- [16] Yi Zhou, Guillermo Gallego, and Shaojie Shen. *Event-based Stereo Visual Odometry*. 2021. arXiv: [2007.15548 \[cs.CV\]](https://arxiv.org/abs/2007.15548).
- [17] V. Usenko, N. Demmel, D. Schubert, J. Stueckler, and D. Cremers. "Visual-Inertial Mapping with Non-Linear Factor Recovery". In: *IEEE Robotics and Automation Letters (RA-L) Int. Conference on Intelligent Robotics and Automation (ICRA) 5.2* (2020), pp. 422–429.
- [18] Tong Qin, Peiliang Li, and Shaojie Shen. "VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator". In: *IEEE Transactions on Robotics* 34.4 (2018), pp. 1004–1020.
- [19] Tong Qin, Jie Pan, Shaozu Cao, and Shaojie Shen. *A General Optimization-based Framework for Local Odometry Estimation with Multiple Sensors*. 2019. eprint: [arXiv:1901.03638](https://arxiv.org/abs/1901.03638).
- [20] Tong Qin, Shaozu Cao, Jie Pan, and Shaojie Shen. *A General Optimization-based Framework for Global Pose Estimation with Multiple Sensors*. 2019. eprint: [arXiv:1901.03642](https://arxiv.org/abs/1901.03642).