SuperEvent: Cross-Modal Learning of Event-based Keypoint Detection for SLAM Supplementary Material

Yannick Burkhardt[1,2,3] Simon Schaefer[1,2,3] Stefan Leutenegger[1,2,3]

[1] Technical University of Munich,
{yannick.burkhardt, simon.k.schaefer}@tum.de
[2] ETH Zürich, lestefan@ethz.ch
[3] Munich Center for Machine Learning (MCML)

1. Ablation Studies

To confirm the effectiveness of our approach, we conduct ablation studies regarding the training data generation process *temporal matching*, and our proposed event representation MCTS. Additionally, we report the performance of our investigated network architectures.

1.1. Training Data

We compare our proposed temporal matching approach with the state-of-the-art method homographic adaptation introduced for frame-based keypoint detection and description [5] in Table 1. SuperEvent trained with temporal matching data exclusively outperforms models trained with homographic adaptation data. Also, combining both approaches results does not improve model performance.

Temporal matching employs real data exclusively without distortions in the event representation due to augmentations. We suspect that this advantage improves the model's event data comprehension.

1.2. Input Representation

Next, we compare our MCTS representation to time surfaces [15] and its variant Tencode [13] in Table 2. As shown in [13], with $\Delta t = 0.01\,\mathrm{s}$, the Tencode model achieves superior performance to the one using time surfaces. However, since Tencode is also used as a single channel tensor, an MCTS with a single time window size (but two channels) strictly separates polarities, thereby pro-

Table 1. Pose estimation after training SuperEvent with temporal matching data, homographic adaptation data, and samples from both methods.

Training Data Generation Method	Pose Estimation AUC in %					
	Event Camera Dataset [22]			Event-aided Direct Sparse Odm.[11]		
	@5°	@10°	@20°	@5°	@10°	@20°
Homographic adaptation	17.2	24.3	31.0	12.3	21.0	31.5
Temporal matching (ours)	**22.7**	**35.8**	**46.7**	**15.2**	**26.4**	**40.1**
Homographic adaptation + temp. matching	18.5	28.1	37.1	13.1	22.0	33.0

Table 2. Pose estimation with different time surface variants as input representations. The index number of the Multi Channel Time Surfaces (MCTS) indicates the number of time window sizes Δt.

Input		Pose Estimation AUC in %					
		Event Camera Dataset [22]			Event-aided Direct Sparse Odometry [11]		
Representation	Channels	@5°	@10°	@20°	@5°	@10°	@20°
Time Surface [15]	1	13.8	21.4	29.3	13.0	22.5	34.1
Tencode (gray) [13]	1	19.5	28.7	36.9	13.9	23.7	36.3
MCTS_1 (ours)	2	20.3	30.0	38.7	14.1	24.4	37.0
MCTS_5 (ours)	10	**22.7**	**35.8**	**46.7**	**15.2**	**26.4**	**40.1**

Table 3. Network architecture ablation study on pose estimation on the Event Camera dataset [22]. Every backbone layer reduces the spatial dimensions by half (except for [3]).
[1] Architecture similar to SuperPoint [5]
[2] Architecture similar to DISK [28]
[3] Architecture similar to SiLK [10] (no spatial reduction in backbone)
[4] SuperEvent
[5] Backbone similar to [9]
[6] Other investigated architectures

	Backbone			Descriptor	Loss		Pose Estimation AUC in %		
	Blocks	Layers	Channels	Resolution	Detector	Descriptor	@5°	@10°	@20°
[1]	VGG	3	32, 64, 128	8x8 grid	Cross-Entropy	dot product	20.2	31.7	42.2
[2]	VGG	3	32, 64, 128	pixelwise	Focal loss	Cycle-Consistency	20.8	31.0	40.7
[3]	VGG	3	32, 64, 128	pixelwise	Focal loss	Cycle-Consistency	18.5	25.5	31.3
[4]	MaxVit	3	32, 64, 128	8x8 grid	Cross-Entropy	dot product	**22.7**	**35.8**	**46.7**
[5]	MaxVit	4	32, 64, 128, 256	8x8 grid	Cross-Entropy	dot product	22.4	33.9	43.8
[6]	MaxVit	3	32, 64, 128	pixelwise	Focal loss	Cycle-Consistency	20.5	29.7	38.0
[6]	MaxVit	3	64, 128, 256	pixelwise	Focal loss	Cycle-Consistency	21.0	30.3	38.6
[6]	MaxVit	4	32, 64, 128, 256	pixelwise	Focal loss	Cycle-Consistency	20.3	30.5	40.3
[6]	MaxVit	4	64, 128, 256, 512	pixelwise	Focal loss	Cycle-Consistency	20.2	29.4	38.2
[6]	MaxVit	5	32, 64, 128, 256, 512	pixelwise	Focal loss	Cycle-Consistency	15.2	23.0	31.3
[6]	MaxVit	5	64, 128, 256, 512, 1024	pixelwise	Focal loss	Cycle-Consistency	17.5	26.4	35.6

viding the model with additional information improving the performance further. Finally, more time windows increase the model's robustness to fast or slow scene motions. Therefore, the 10-channel MCTS$_5$ with $\Delta t_{\{1,\ldots,N\}} = \{0.001\,\mathrm{s}, 0.003\,\mathrm{s}, 0.01\,\mathrm{s}, 0.03\,\mathrm{s}, 0.1\,\mathrm{s}\}$ enables the model to outperform all other variants.

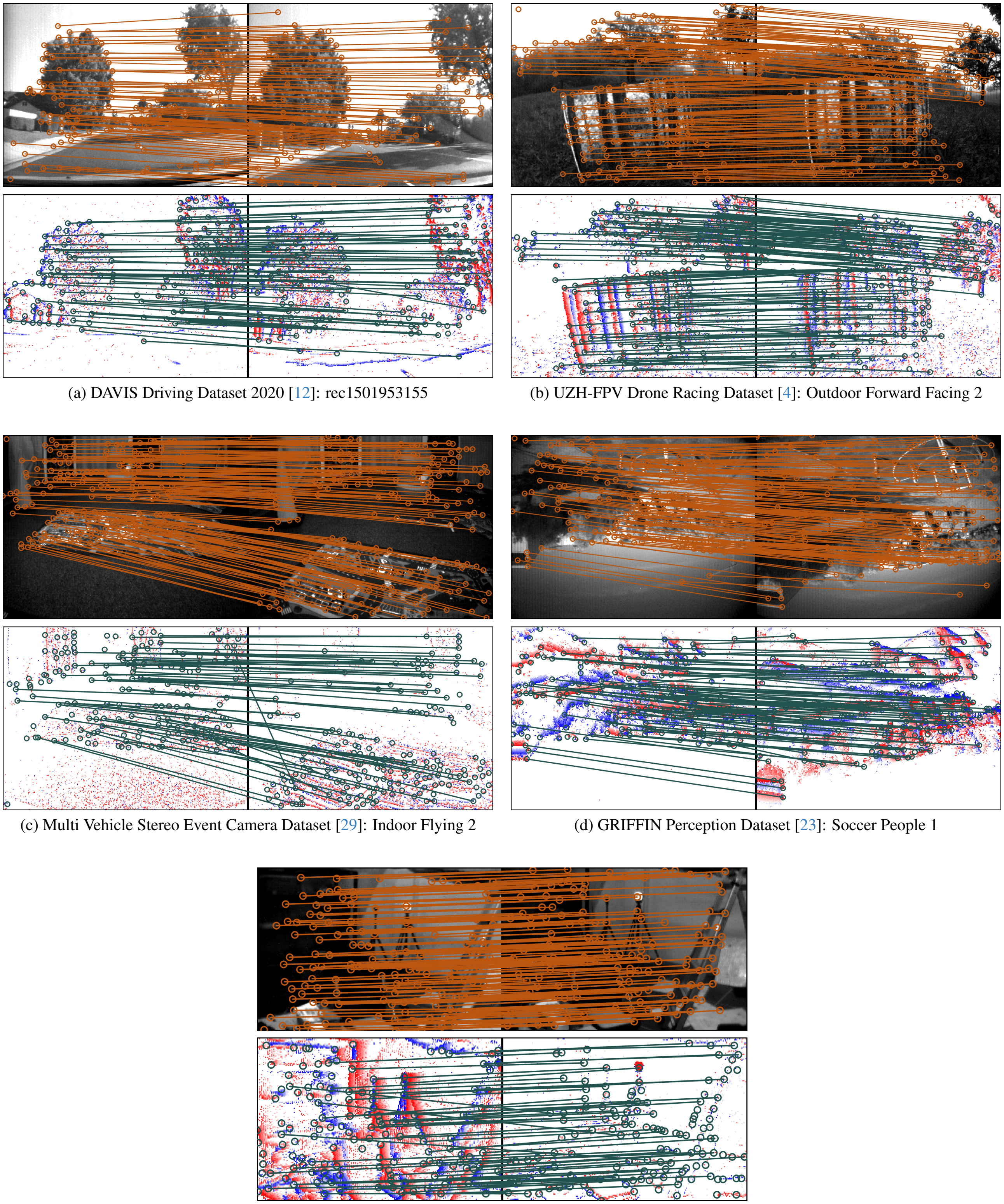
1.3. Network Architecture

We compare various combinations of network architectures from the literature [5, 9, 10]. We investigate two backbone configurations, namely VGG [25] and MaxVit [27], and their hyperparameters *number of layers in backbone* and *output channels per layer in backbone*. Additionally, we investigated if a descriptor prediction on pixel-level as in [10, 28] performs better than the 8-grid interpolation from [5]. For the pixelwise descriptor approach, we employ Focal loss [20] to train the detector head and the Cycle-Consistency loss [10, 28] for the descriptor head.

2. Examples of Network Predictions and Pseudo-labels

Figure 1 shows training samples of SuperEvent generated by temporal matching of gray-scale frames. Due to the modality change, SuperEvent does not learn to exactly match the pseudo-labels, but partially detects and matches different keypoints, while still yielding similar patterns.

Matched keypoints from SuperEvent on unseen sequences are shown in Figure 2. These sequences are held out during training, showing the generalization ability of SuperEvent.



(a) DAVIS Driving Dataset 2020 [12]: rec1501953155

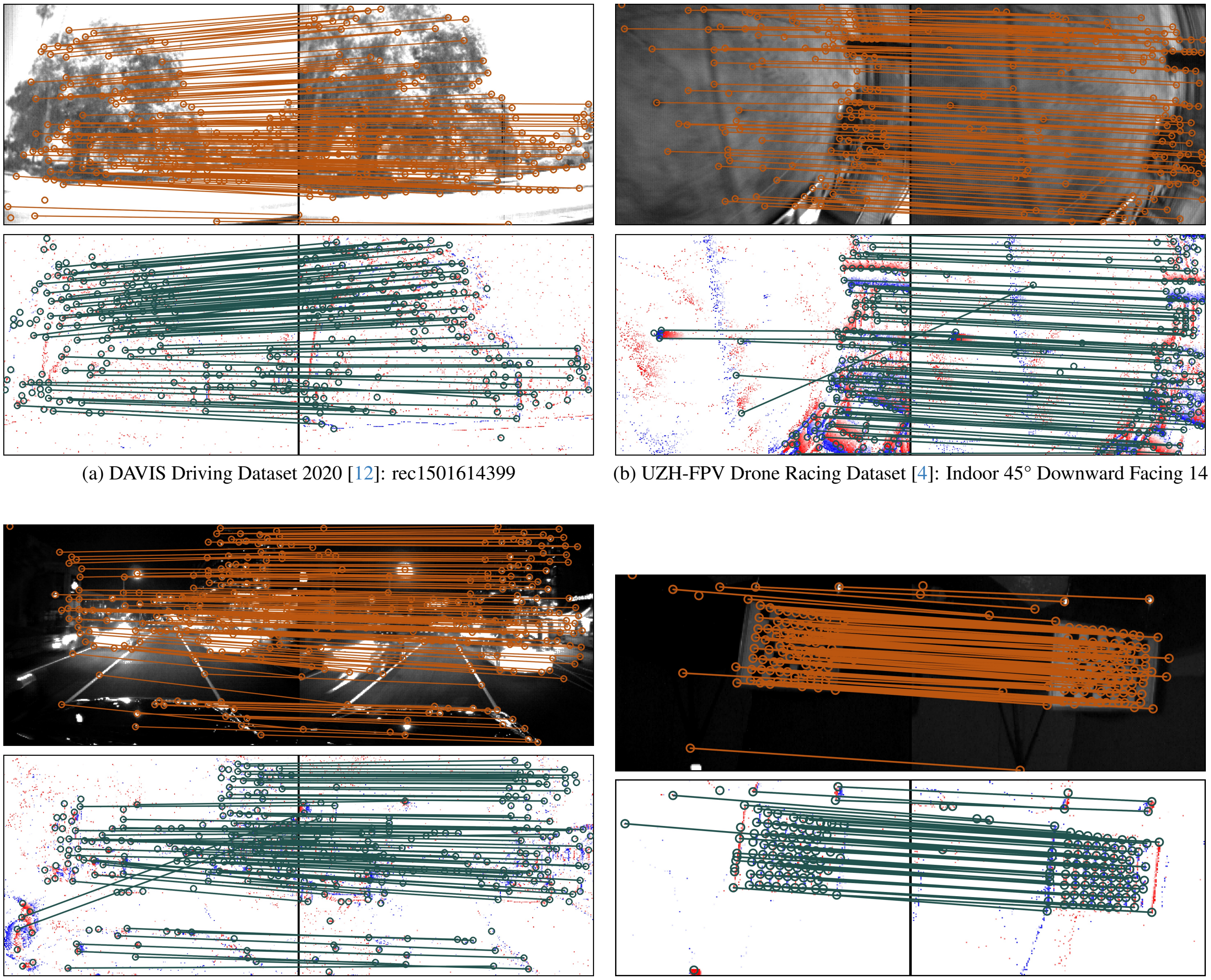
(b) UZH-FPV Drone Racing Dataset [4]: Outdoor Forward Facing 2

(c) Multi Vehicle Stereo Event Camera Dataset [29]: Indoor Flying 2

(d) GRIFFIN Perception Dataset [23]: Soccer People 1

(e) Vision for Visibility Dataset [16]: Indoor Global Aggressive

Figure 1. Examples of the training data for temporal matching. Top (orange): pseudo-labels generated by SuperPoint [5] + SuperGlue [24]; bottom (green): predictions of SuperEvent after training.



(a) DAVIS Driving Dataset 2020 [12]: rec1501614399

(b) UZH-FPV Drone Racing Dataset [4]: Indoor 45° Downward Facing 14

(c) Multi Vehicle Stereo Event Camera Dataset [29]: Outdoor Night Drive 1

(d) Vision for Visibility Dataset [16]: Indoor Varying Robust

Figure 2. Examples of predictions and pseudo-labels of data not used for training. Top (orange): pseudo-labels generated by SuperPoint [5] + SuperGlue [24]; bottom (green): predictions of SuperEvent.

3. Event-based versus Frame-based Keypoint Matching

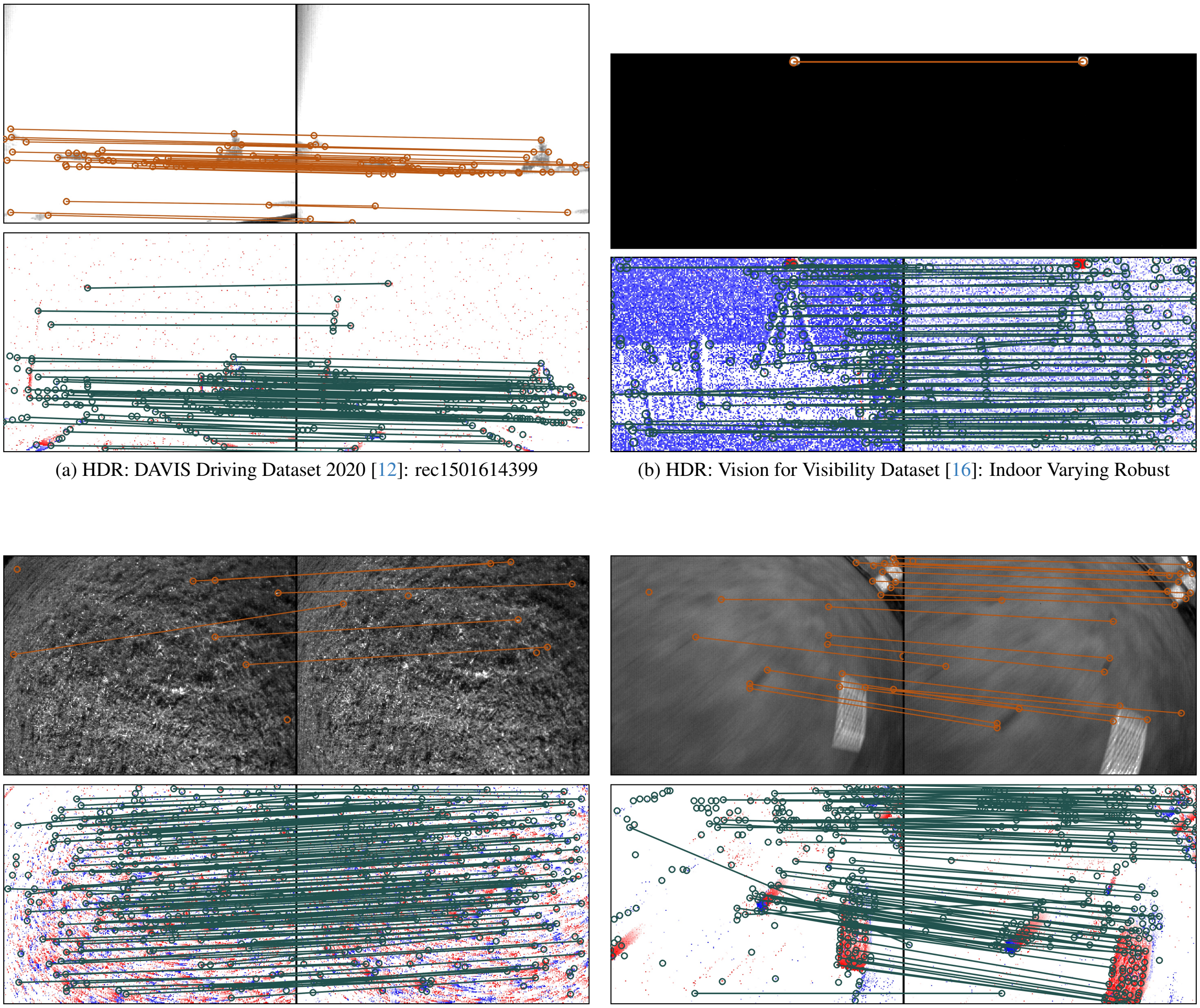
Since SuperEvent is not explicitly trained in scenarios where the quality of frame cameras degrades, such as fast scene motion and high dynamic range (HDR), we demonstrate its generalization ability by comparing SuperEvent's event-based keypoint correspondences to related frame-based results.

3.1. Qualitative Comparison for Fast Motion and HDR

Figure 3 shows matches of the same scenes (not used for training) generated by SuperPoint [5] + SuperGlue [24] on the frames and by SuperEvent and brute-force matching on the event stream. While the quality of the frames under fast scene motion and HDR degrades, the event stream suffers less under these conditions, resulting in better and more equally distributed keypoint matches when using SuperEvent.

3.2. Comparison in SLAM Downstream Task

We quantitatively compare the SLAM results from plain frame-based OKVIS2 [17] and our modified version after replacing the frame-based BRISK [18] detector with SuperEvent. Since event data degrades less than frames in such conditions, SuperEvent's predictions improve OKVIS2's estimations. Choosing a higher processing rate further boosts the performance: the visual overlap between MCTSs rises, and the effects of motion dependence are mitigated.



(a) HDR: DAVIS Driving Dataset 2020 [12]: rec1501614399

(b) HDR: Vision for Visibility Dataset [16]: Indoor Varying Robust

(c) Fast motion: UZH-FPV Drone Racing Dataset [4]: Outdoor 45° Downward Facing 1

(d) Fast motion: UZH-FPV Drone Racing Dataset [4]: Indoor 45° Downward Facing 14

Figure 3. Frame-based SuperPoint [5] + SuperGlue [24] (top), event-based SuperEvent (ours, bottom) on unseen sequences, showing the superior matching capabilities of SuperEvent by leveraging the higher quality of event data for scenes with HDR or fast scene motion.

Table 4. Results on TUM-VIE [14] sequences with fast motion (*mocap-shake*) and low light (*floor2-dark*). ATE and RPE in cm; RPE for consecutive frames. Results marked * are not representative due to discontinuous ground truth.

OKVIS2 [17]	**+ SuperEvent (ours)**				frame-based	
	20 Hz		40 Hz		20 Hz	
Sequence	ATE	RPE	ATE	RPE	ATE	RPE
mocap-shake	43.71	0.55	**29.14**	**0.26**	50.83	1.03
mocap-shake2	43.75	0.80	**27.37**	**0.46**	66.29	1.39
floor2-dark	9.58	2.51*	**9.37**	**1.23***	failed	failed

4. Pose Estimation Experiment

In this section, we explain the details of the keypoint-based pose estimation benchmark and justify why we chose this method to evaluate SuperEvent.

4.1. Benchmark Design

Pose Estimation requires reliable keypoint detection and matching and is therefore a common baseline for frame-based keypoint detectors [10, 24, 26, 28]. It also indicates the approaches' usability for downstream applications such as Visual Odometry, Simultaneous Localization and Mapping (SLAM), and Structure-from-Motion (SfM) that usually rely on keypoint-based pose estimation. Frame-based approaches are evaluated on datasets such as ScanNet [3] and MegaDepth [19] containing various images of the same scene with the associated ground truth camera poses, allowing for a straightforward pose estimation evaluation. However, event datasets are usually temporally continuous sequences because of the sensor's asynchronous nature. This raises the question, at which timestamps the pose estimation is evaluated. We define the evaluation benchmark as follows:

- At each time step t_i with available ground truth data, the ground truth camera orientation must change by the maximal rotation change angle $c_{\Delta \mathrm{r},\mathrm{max}}$ within the maximal time difference $c_{\Delta \tau}$.
- We find equally n distributed rotation changes in this time interval to evaluate the pose estimation.

In this experiment, we choose $c_{\Delta \mathrm{r},\mathrm{max}} = 45\,°$ $c_{\Delta \tau} = 2\,\mathrm{s}$ and $n_{\Delta \mathrm{r}} = 45$ to test various levels of difficulty while reducing the amount of samples without visual overlap. Our practical implementation executes the following steps for each sequence:

1. For all timestamps with associated ground truth pose measurements (usually between 100 and 200 Hz), we generate keypoint and descriptor predictions. For tracking approaches, we assign a track ID as a scalar descriptor. Matching the same track ID reproduces the tracking result.
2. Next, we iteratively check for each ground truth sample if any of the subsequent ground truth samples within $c_{\Delta \tau} = 2\,\mathrm{s}$ yields an rotation change of at least $c_{\Delta \mathrm{r},\mathrm{max}} = 45\,°\}$. If this condition cannot be fulfilled, we skip the respective sample.
3. For samples with sufficient rotation change within $c_{\Delta \mathrm{r},\mathrm{max}}$, we find the first subsequent ground truth rotation values that surpass the equally distributed rotation changes $\frac{c_{\Delta \mathrm{r},\mathrm{max}}}{n_{\Delta \mathrm{r}}} \cdot \{1, 2, \dots, n_{\Delta \mathrm{r}}\} = \{1\,°, 2\,°, \dots, 45\,°\}$.
4. We match the keypoint descriptors for these selected prediction samples with associated ground truth measurements. The camera pose is estimated based on the resulting keypoint pairs and considering the lens distortion (unless the approach to evaluate already required a rectified event representation as input). We calculate the rotation difference and its angle of the axis-angle representation.
5. Having evaluated samples of the dataset sequences, we report the area-under-curve (AUC) for different thresholds as in [26].

We evaluate SuperEvent on the following sequences, omitting the ones without sufficient rotation changes as well as calibration sequences.

Event Camera Dataset [22]:

- boxes_6dof
- boxes_rotation
- poster_6dof
- poster_rotation
- shapes_6dof
- shapes_rotation

Event-aided Direct Sparse Odometry [11]:

- peanuts_dark
- peanuts_light
- rocket_earth_light
- rocket_earth_dark
- ziggy_and_fuzz
- ziggy_and_fuzz_hdr
- peanuts_running
- all_characters

4.2. Why Not Homography Estimation?

Some of the existing approaches are benchmarked on homography estimation of planar scenes [1, 2, 8, 13, 21]. However, the commonly used HVGA ATIS Corner dataset [21] contains neither ground truth poses nor homography measurements. Therefore, the authors compare the mean reprojection error (MRE) of their method's detected points warped by the estimated homography. But without comparing it to any ground truth, in general, it cannot be guaranteed that the estimated homography is (close

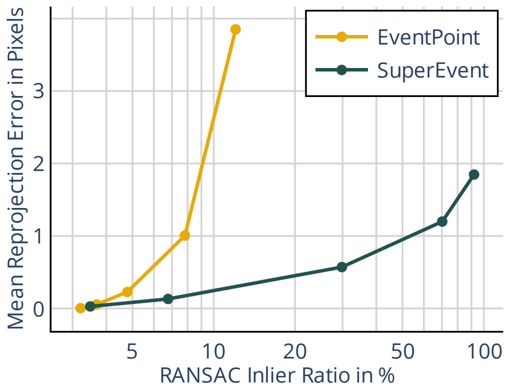


Figure 4. Homography estimation on HVGA ATIS [21] matching 400 keypoints per sample, $\Delta t = 50\,\mathrm{ms}$, and *RANSAC reprojection error thresholds* of $\{0.1, 0.3, 1, 3, 10\}$ pixels.

Table 5. RPE scores in cm for consecutive samples at 20 Hz of OKVIS2 [17] + SuperEvent on the TUM-VIE *mocap* sequences.

1d-trans	3d-trans	3dof	desk	desk2
0.07	0.14	0.10	1.39	1.25

to) correct. E.g., in most cases, a nonsensical estimate that matches four random keypoints can achieve the optimal score of 0 since this is the minimum number of point correspondences to estimate the homography; and there will not be any outliers that negatively influence the score.

As long as there are always sufficiently many keypoints detected, this evaluation procedure might still be sensible for approaches that rely on basic nearest neighbor matching in pixel-space because some wrong matches do not lead to large errors. However, for approaches that rely on descriptor matching [1, 2, 21], only a few outliers with large errors have a serious negative impact on the reported score. Most downstream applications therefore employ outlier filtering, such as Random Sample Consensus (RANSAC) [6]. Also, the approaches relying on descriptor matching [8, 13] employ RANSAC as their homography estimation benchmark. The RANSAC algorithm rejects outliers with a reprojection error greater than a pre-defined threshold c_{R}. Thus, this threshold is an upper bound to the estimated homography reprojection error since all keypoints with larger errors are filtered out. Thereby, the MRE score can be arbitrarily reduced by choosing a smaller c_{RE}, making it inappropriate for performance benchmarking.

This general problem applies not only to RANSAC but to all approaches relying on some form of outlier removal as a post-processing step. Outliers, of course, are never completely avoidable, and it is a common procedure to remove them in downstream applications. Therefore, we decided to reproduce the frame-based benchmark of estimating the camera pose change in datasets with ground truth camera pose measurements – producing meaningful results for approaches with outlier removal.

We illustrate this issue in Figure 4 where we plot the reprojection error after homography estimation for different *RANSAC reprojection error thresholds* over the ratio of matches classified as inliers: SuperEvent achieves similar reprojection errors as EventPoint [13] with RANSAC leading to far fewer outliers.

5. Stereo Event-Visual Intertial SLAM Experiment

Lastly, we visualize 2D projections of the trajectories estimated by SuperEvent integrated into OKVIS2 [17] yielding the reported results. In addition to the ATE reported in the main paper, we also report the RPE results of our method for completeness.

Table 6. RPE scores in cm for consecutive samples at 20 Hz of OKVIS2 [17] + SuperEvent on the VECtor large-scale sequences.

corr.- dolly	corr.- walk	units- dolly	units- scooter	school- dolly	school- scooter
16.72	14.50	23.10	42.40	21.66	45.78

5.1. TUM-VIE Small-scale Sequences

Figure 5 shows the trajectories from SuperEvent + OKVIS2 on the TUM-VIE [14] *mocap*-sequences. Since OKVIS2 is non-deterministic, we process every sequence 5 times and select the trajectory with median error. We report the RPE scores in Table 5.

5.2. TUM-VIE Large-scale Sequences

The effect of loop closure on the trajectory estimation of OKVIS2 + SuperEvent can be seen in Figure 6. The loop closure is reliably detected on all 4 *loop-floor* sequences of the TUM-VIE dataset. Since the ground truth is not continuous, we do not report RPE scores for these sequences, as they lack interpretive value and are not comparable.

5.3. VECtor Large-scale Sequences

Figure 7 shows the trajectories from SuperEvent + OKVIS2 on the VECtor [7] large-scale sequences. We report RPE scores in Table 6.

References

[1] Philippe Chiberre, Etienne Perot, Amos Sironi, and Vincent Lepetit. Detecting stable keypoints from events through image gradient prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1387–1394, 2021. 6, 7

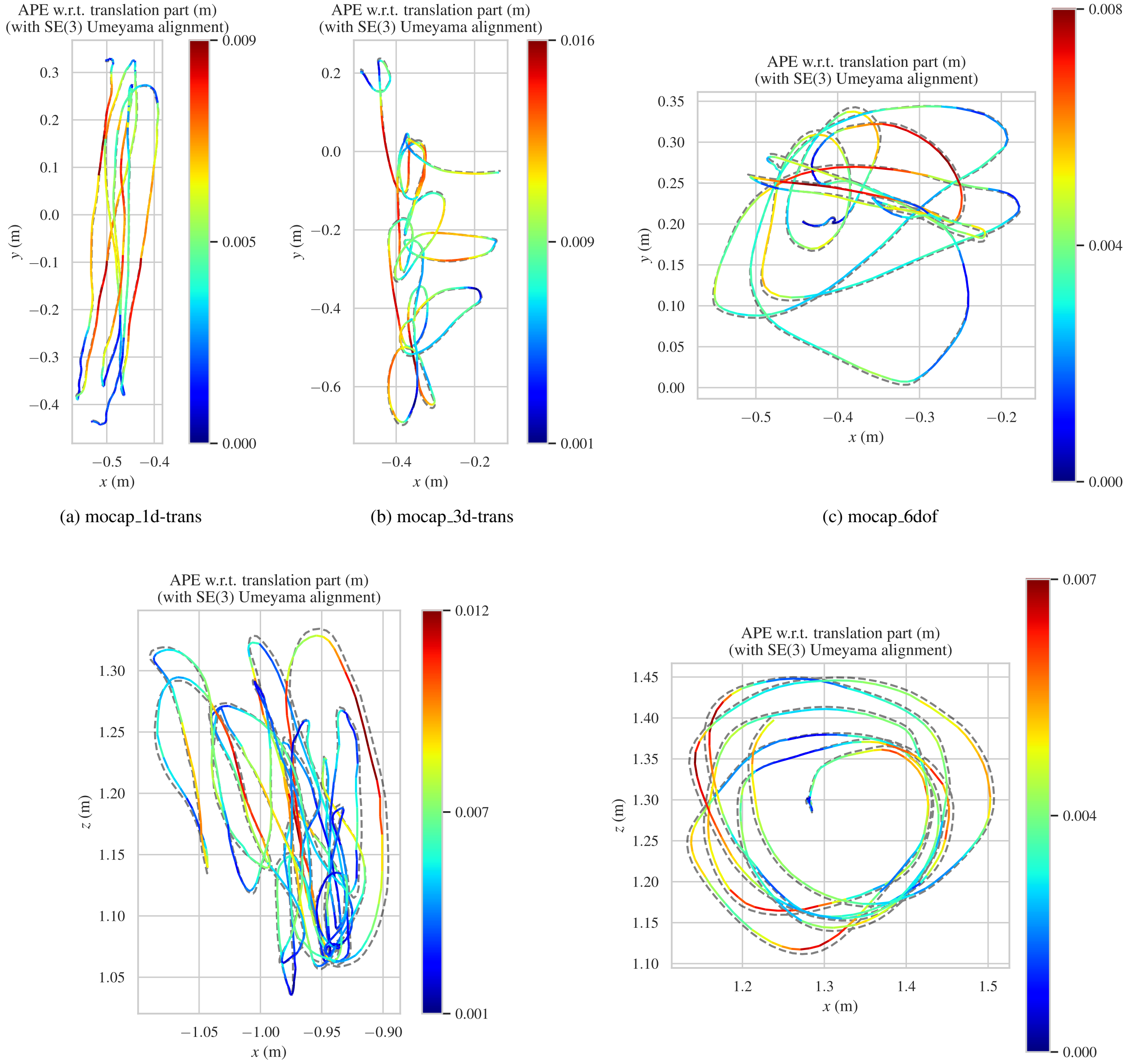
[2] Philippe Chiberre, Etienne Perot, Amos Sironi, and Vincent Lepetit. Long-lived accurate keypoints in event streams. *arXiv preprint arXiv:2209.10385*, 2022. 6, 7

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6

[4] 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 *2019 International Conference on Robotics and Automation (ICRA)*, pages 6713–6719. IEEE, 2019. 3, 4, 5

[5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 1, 2, 3, 4, 5

[6] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to



(a) mocap_1d-trans

(b) mocap_3d-trans

(c) mocap_6dof

(d) mocap_desk

(e) mocap_desk2

Figure 5. OKSVIS2 + SuperEvent's trajectories on the TUM-VIE *mocap* sequences with ground truth trajectory (dashed) and absolute position error of the 3D trajectories shown by the color.

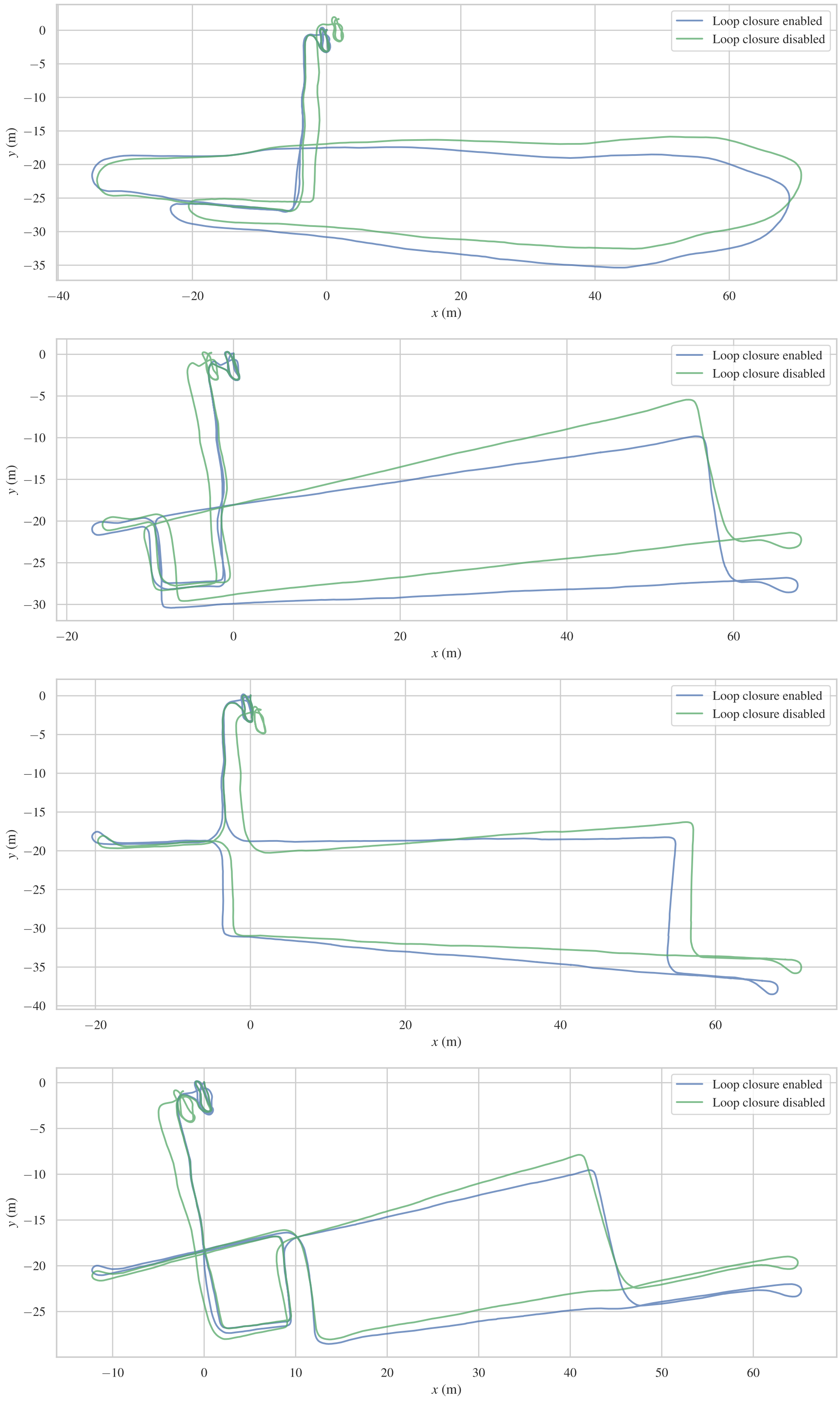
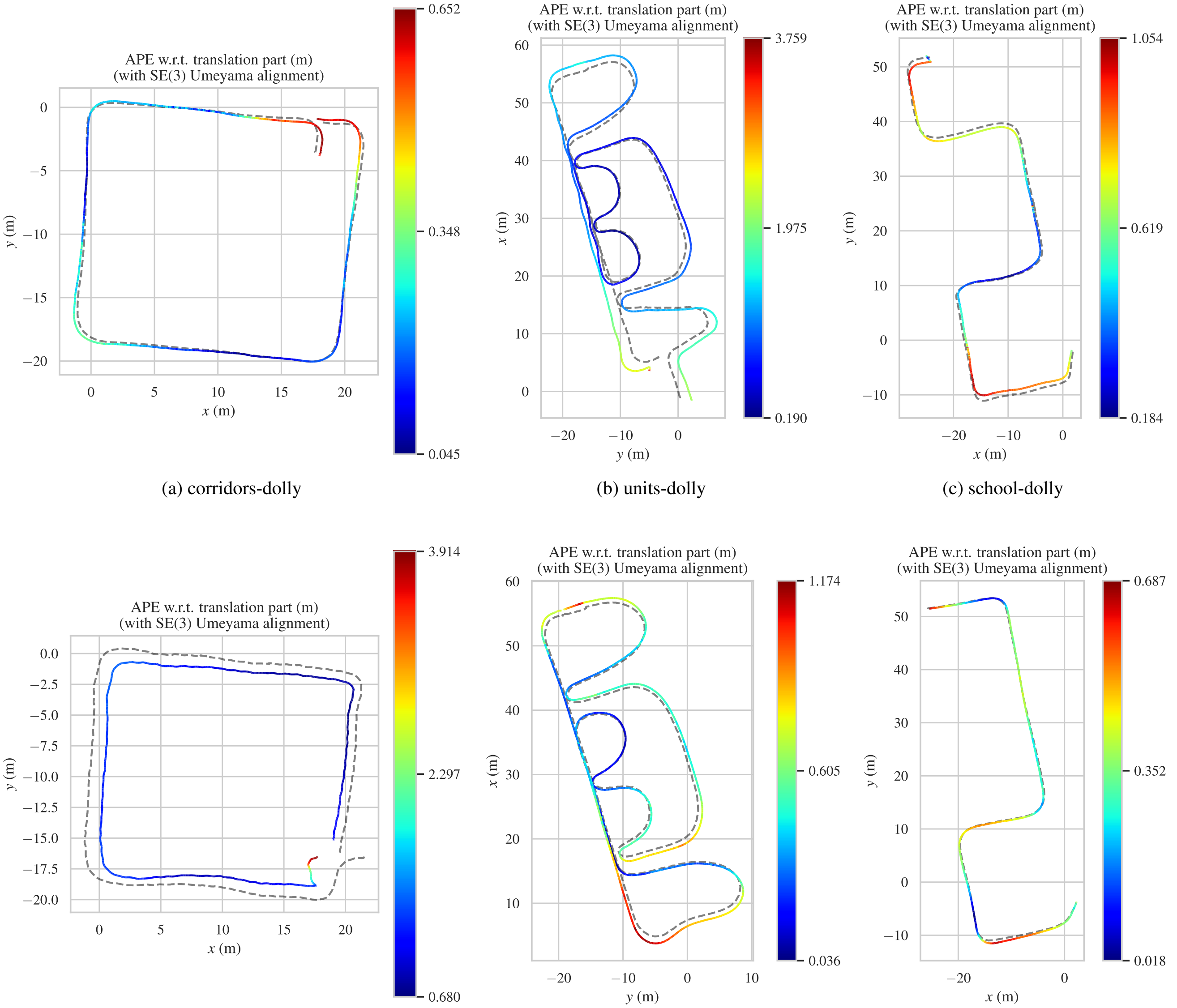


Figure 6. OKSVIS2 + SuperEvent's trajectories on the TUM-VIE loop-floor 0-3 sequences with and without loop closure.



(a) corridors-dolly

(b) units-dolly

(c) school-dolly

(d) corridors-walk

(e) units-scooter

(f) school-scooter

Figure 7. OKSVIS2 + SuperEvent's trajectories on the VECtor large-scale sequences with ground truth trajectory (dashed) and absolute position error of the 3D trajectories shown by the color.

image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 7

[7] Ling Gao, Yuxuan Liang, Jiaqi Yang, Shaoxun Wu, Chenyu Wang, Jiaben Chen, and Laurent Kneip. Vector: A versatile event-centric benchmark for multi-sensor slam. *IEEE Robotics and Automation Letters*, 7(3):8217–8224, 2022. 7

[8] Yuan Gao, Yuqing Zhu, Xinjun Li, Yimin Du, and Tianzhu Zhang. SD2Event: Self-supervised learning of dynamic detectors and contextual descriptors for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3055–3064, 2024. 6, 7

[9] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13884–13893, 2023. 2

[10] Pierre Gleize, Weiyao Wang, and Matt Feiszli. SiLK: Simple learned keypoints. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22442–22451, 2023. 2, 6

[11] Javier Hidalgo-Carrió, Guillermo Gallego, and Davide Scaramuzza. Event-aided direct sparse odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5781–5790, 2022. 1, 6

[12] Yuhuang Hu, Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. DDD20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6. IEEE, 2020. 3, 4, 5

[13] Ze Huang, Li Sun, Cheng Zhao, Song Li, and Songzhi Su. EventPoint: Self-supervised interest point detection and description for event-based camera. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5396–5405, 2023. 1, 6, 7

[14] Simon Klenk, Jason Chui, Nikolaus Demmel, and Daniel Cremers. TUM-VIE: The TUM stereo visual-inertial event dataset. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8601–8608. IEEE, 2021. 5, 7

[15] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E. Shi, and Ryad B. Benosman. HOTS: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1346–1359, 2017. 1

[16] Alex Junho Lee, Younggun Cho, Young-sik Shin, Ayoung Kim, and Hyun Myung. ViViD++: Vision for visibility dataset. *IEEE Robotics and Automation Letters*, 7(3):6282–6289, 2022. 3, 4, 5

[17] Stefan Leutenegger. OKVIS2: Realtime scalable visual-inertial SLAM with loop closure, 2022. 4, 5, 7

[18] Stefan Leutenegger, Margarita Chli, and Roland Y. Siegwart. BRISK: Binary robust invariant scalable keypoints. In *2011 International Conference on Computer Vision*, pages 2548–2555, 2011. 4

[19] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 6

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 2

[21] Jacques Manderscheid, Amos Sironi, Nicolas Bourdis, Davide Migliore, and Vincent Lepetit. Speed invariant time surface for learning to detect corner points with event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10245–10254, 2019. 6, 7

[22] 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. 1, 2, 6

[23] Juan Pablo Rodríguez-Gómez, Raul Tapia, Julio L Paneque, Pedro Grau, Augusto Gómez Eguíluz, Jose Ramiro Martínez-de Dios, and Anibal Ollero. The GRIFFIN perception dataset: Bridging the gap between flapping-wing flight and robotic perception. *IEEE Robotics and Automation Letters*, 6(2):1066–1073, 2021. 3

[24] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 3, 4, 5, 6

[25] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[26] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 6

[27] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MaxViT: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022. 2

[28] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 2, 6

[29] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3 (3):2032–2039, 2018. 3, 4