Supplementary Material

7. Method Details

Clarifications Event Representation Events are quasi-continuous. Equation (2) defines the task of tracking any point from events as determining the time-discrete point observations from the continuous input events. In the first step events are converted to event representations, where each representation has a constant number of events N_e. Section 8 shows exemplary the connection between events and discrete tracking timesteps τ, resulting in a constant tracking frequency, despite a varying event rate. Please note that the tracking frequency is adjustable at test time. In practice, we mostly set τ_t to the ground truth timesteps of an evaluation set.

Description of Event Stacks As frame representation, we use a variation of Mixed-Density event stacks [46] and build T input representations I_t. Let $E_t = \{e_i | t_i \leq \tau_t\}$ be the N_e events directly preceding timestep τ_t. We construct a multi-channel representation by hierarchically binning these events into $C = 10$ channels, denoted as $\{h_c\}_{c=1}^{C}$, where each channel h_c is a spatial histogram of dimensions $H \times W$. The c-th channel aggregates $n_c = \lfloor N_e / 2^{c-1} \rfloor$ events using bilinear interpolation, such that:

- h_1 incorporates all N_s events
- h_c processes $N_s / 2^{c-1}$ events for $c > 1$

where each channel contains the events closest to t_i.

Hyperparameters For a better overview Tab. 6 provides an overview of all hyperparameters of our method introduced in Sec. 3.

Event Generation Model The linear event generation model has been discussed previously (e.g. [17]. To make the paper self-contained, here is a brief introduction. It approximates how events are triggered in event cameras. Starting from the condition that events occur when brightness change reaches a threshold ($\Delta L(\mathbf{x}_k, t_k) = p_k C$), this model uses Taylor's expansion for small time intervals to relate events to the temporal derivative of brightness ($\Delta L t(\mathbf{x}_k, t_k) \approx \frac{p_k C}{\Delta t_k}$). Under constant illumination, this can be further linearized to $\Delta L \approx -\nabla L \cdot v \Delta t$, showing that events are fundamentally triggered by brightness gradients (edges) moving across the image plane. The rate of event generation depends on the relationship between edge orientation and motion direction, with perpendicular motion producing the highest event rate.

Events under Time Inversion. According to the linearized event generation model (LEGM) [17] an event e_k is generated when the dot product between per-pixel optical flow v and the image gradient ∇L exceeds the threshold C:

$$e_k \in E_t \iff -p_k \nabla L(\mathbf{x}_k, \tau_k) \cdot v(\mathbf{x}_k, \tau_k) \delta\tau_k \approx C \quad (8)$$

where $\delta\tau_k$ is the time since the last event at the same pixel.

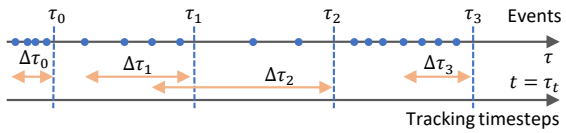

Figure 9. Asynchronous events are converted into temporally equidistant frame representations at τ_t, each created from the last N_e events.

Parameter	Variable	Value
window length	w	8
feature size	d	128
bin number	B	10
stride	T_s	4
refinement steps (train)	M	4
refinement steps (eval)	M	6
feature scales	S	4

Table 6. *Hyperparameters*. An overview of variables that were introduced in Sec. 3 and their specific values.

Next, consider how the events E_t change when the motion changes, for example, induced by a time inversion $\tilde{\tau} \doteq 2\bar{\tau}_t - \tau$, with $\bar{\tau}_t = \frac{\tau_t + \tau_t - \Delta\tau_t}{2}$ is the interval midpoint. Due to the chain rule, the optical flow becomes $\tilde{v}(\mathbf{x}, \tau) = -v(\mathbf{x}, 2\bar{\tau}_t - \tau)$, and the gradient becomes $\nabla\tilde{L}(\mathbf{x}, \tau) = \nabla L(\mathbf{x}, 2\bar{\tau}_t - \tau)$. Under this change of variables, we describe what the new events $\tilde{E}_t$ look like. Specifically, if $e_k \in E_t$, then $\tilde{e}_k = (\mathbf{x}_k, 2\bar{\tau}_t - \tau_k, -p_k) \in \tilde{E}_t$ since

$$\begin{aligned} &- \tilde{p}_k \nabla\tilde{L}(\tilde{\mathbf{x}}_k, \tilde{\tau}_k) \cdot \tilde{v}(\tilde{\mathbf{x}}_k, \tilde{\tau}_k) \delta\tilde{\tau}_k \\ &= -p_k \nabla L(\mathbf{x}_k, \tau_k) \cdot v(\mathbf{x}_k, \tau_k) \delta\tau_k \overset{(8)}{\approx} C. \end{aligned} \quad (9)$$

The equality is satisfied assuming the time since the last event is similar under time inversion ($\delta\tilde{\tau}_k \approx \delta\tau_k$). Simple inspection shows that the events E_t and $\tilde{E}_t$ are different, and, as a result, corresponding descriptors D_t^s and $\tilde{D}_t^{w-s+1}$ are different (note $w - s + 1$ is the inverted index).

8. Data and Evaluation Details

8.1. Ground truth generation for the E2D2 Fidget Spinner Sequence

The ground truth tracks used for evaluation on the E2D2 fidget spinner sequence were calculated from simple geometric knowledge. The midpoint of the spinner is constant. The wheel itself is fully facing the camera, describing perfect circular motions. Therefore, we can calculate the positions of each point on the fidget spinner with an estimate of the angular velocity of the wheel. The angular velocity is estimated as follows: First, we create event histograms with a fixed number of 20,000 events at 1000 Hz (simply counting

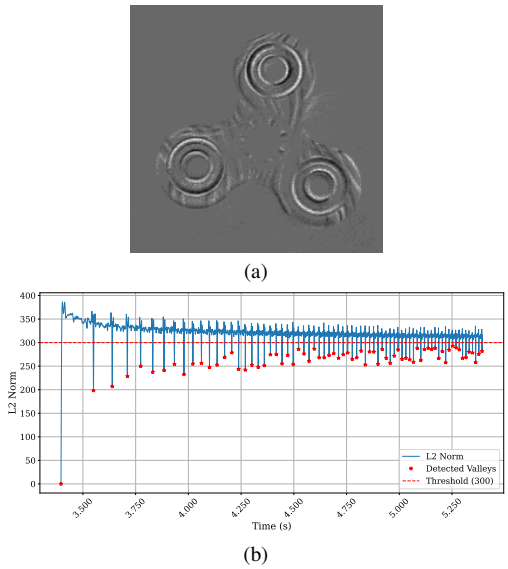

Figure 10. *Ground truth for E2D2 fidget spinner sequence*. (a) Example of a 2D event histogram that is built at 1000Hz. (b) time series of L2 norms wrt. to the first frame. Red star points are local minima, where the spinner completed another third revolution.

positive and negative events within the event batch), as seen in Figure 10 (a). Then we calculate the 1D time series of the L2-norm between each frame and the initial frame, visualized in Fig. 10 (b). The local minima are the times when the wheel completed a third revolution (due to the three-lobed shape of the fidget spinner). We assume the angular velocity to be constant between two third-revolution-timestamps. As shown in Fig. 10 (b), the spinner gets progressively faster, increasing tracking difficulty.

8.2. Examples of the EventKubric Dataset

Figure 11 visualizes the data generation explained in Sec. 4. Figure 13 shows a few examples of the EventKubric dataset. The full scene knowledge is available as annotations, which can be useful for tasks beyond point tracking.

9. Further Experiments and Detailed Results

9.1. Task 2: Feature Tracking - Extended Results

Table 8 provides full results for the EDS & EC dataset. Figure 15 shows additional comparisons.

9.2. Results EVIMO2

Figure 14 shows prediction results for EVIMO2

9.3. Feature Independence Experiment.

We examine the effect of our contrastive loss on the learned features with an experiment shown in Fig. 12. We track the same 3 points on a 2D pattern with two orthogonal

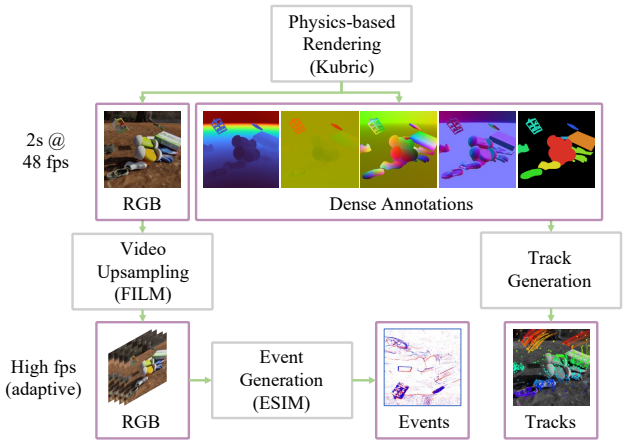

Figure 11. *Data Generation Pipeline*. The PBR tool Kubric renders 2s RGB videos, which are adaptively upsampled to generate events from it. The dense ground truth provided by Kubric is used for point track generation.

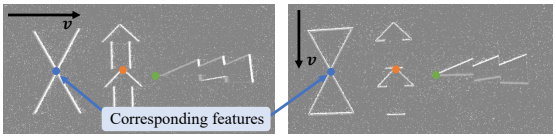

Figure 12. *Setup of the motion robustness experiment*. The same pattern is recorded two times in perpendicular directions at the same key points of the pattern. The same points under different motion directions should ideally have similar descriptors.

Method	$\mathcal{C}_{\text{intra}} \uparrow$	$\mathcal{C}_{\text{inter}} \uparrow$	Δ
Frames	0.836	0.804	**0.032**
Events without FA-loss	0.776	0.399	0.377
Events with FA-loss	**0.954**	**0.887**	0.067

Table 7. *Measuring feature independence*. The intra- and inter-cluster cosine similarity of tracking the same points in different sequences.

camera motions and analyze the corresponding descriptors $d^i_{t,\text{dir}}$ at the end of the window with point index i and dir $\in$ {horizontal, vertical}. We then measure the cosine similarity between descriptors at the trajectory start, and descriptors along the same trajectory with $\mathcal{C}_{\text{intra}} = \sum_{t,\text{dir},i} \cos_{\text{sim}}(d^i_{0,\text{dir}}, d^i_{t,\text{dir}})$, called *intra-cluster*, and along trajectories with *different motions directions* e.g. $\mathcal{C}_{\text{inter}} = \sum_{t,i} \cos_{\text{sim}}(d^i_{0,\text{horizontal}}, d^i_{t,\text{vertical}})$, called *inter-cluster*. Table 7 shows results for three methods: our model, an ablation model trained without our loss, and a frame-based baseline. While the model in the motion-independent frame domain has very similar inter- and intra-cluster similarities, the ablation model shows a similarity gap of 0.38 between $\mathcal{C}_{\text{intra}}$ and $\mathcal{C}_{\text{inter}}$. In comparison, this gap is closed, when training with our contrastive loss.

Method	Frames	Average		Peanuts Light		Rocket Earth*		Ziggy Arena		Peanuts Running	
		FA↑	EA↑	FA↑	EA↑	FA↑	EA↑	FA↑	EA↑	FA↑	EA↑
EKLT [21]	✓	0.325	0.325	0.284	0.260	0.425	0.175	0.419	0.231	0.171	0.153
DDFT [44]	✓	0.576	0.472	0.447	0.420	0.648	0.291	0.748	0.746	0.460	0.428
FE-TAP [38]	✓	0.676	0.589	**0.549**	**0.517**	0.538	0.246	**0.849**	**0.844**	**0.769**	**0.749**
ICP [32]	✗	0.060	0.040	0.050	0.044	0.103	0.045	0.043	0.039	0.043	0.028
EM-ICP [63]	✗	0.161	0.120	0.084	0.077	0.298	0.158	0.153	0.149	0.108	0.095
HASTE [3]	✗	0.096	0.161	0.086	0.076	0.162	0.085	0.082	0.057	0.054	0.033
DDFT E2VID [44]	✗	0.589	0.495	–	–	–	–	–	–	–	–
ETAP w\o FA-loss (Ours)	✗	0.698	0.599	0.538	0.508	0.676	0.336	0.842	0.841	0.736	0.713
ETAP (Ours)	✗	0.705	0.598	0.529	0.5	0.705	0.336	0.839	0.838	0.746	0.717

Method	Frames	Average		shapes_trans		shapes_rot		shapes_6dof		boxes_trans		boxes_rot	
		FA↑	EA↑	FA↑	EA↑	FA↑	EA↑	FA↑	EA↑	FA↑	EA↑	FA↑	EA↑
EKLT [21]	✓	0.811	0.775	0.839	0.740	0.833	0.806	0.817	0.696	0.682	0.644	0.883	0.865
DDFT [44]	✓	0.825	0.818	0.861	0.865	0.797	0.793	0.899	0.882	0.872	0.869	0.695	0.691
FE-TAP [38]	✓	0.844	0.838	**0.931**	**0.929**	0.815	0.813	0.879	0.860	0.731	0.728	0.862	0.861
ICP [32]	✗	0.256	0.245	0.307	0.306	0.341	0.339	0.169	0.129	0.268	0.261	0.191	0.188
EM-ICP [63]	✗	0.337	0.334	0.403	0.402	0.320	0.320	0.248	0.242	0.355	0.354	0.356	0.349
HASTE [3]	✗	0.442	0.427	0.589	0.564	0.613	0.582	0.133	0.043	0.382	0.368	0.492	0.447
DDFT E2VID [44]	✗	0.794	0.786	–	–	–	–	–	–	–	–	–	–
ETAP w\o FA-loss (Ours)	✗	0.885	0.879	0.904	0.902	0.868	0.867	0.91	0.891	0.879	0.877	0.866	0.863
ETAP (Ours)	✗	0.888	0.883	0.91	0.904	0.867	0.865	0.904	0.886	0.866	0.864	0.896	0.893

Table 8. Detailed performance comparison of tracking methods on the EDS (top) and EC (bottom) datasets.

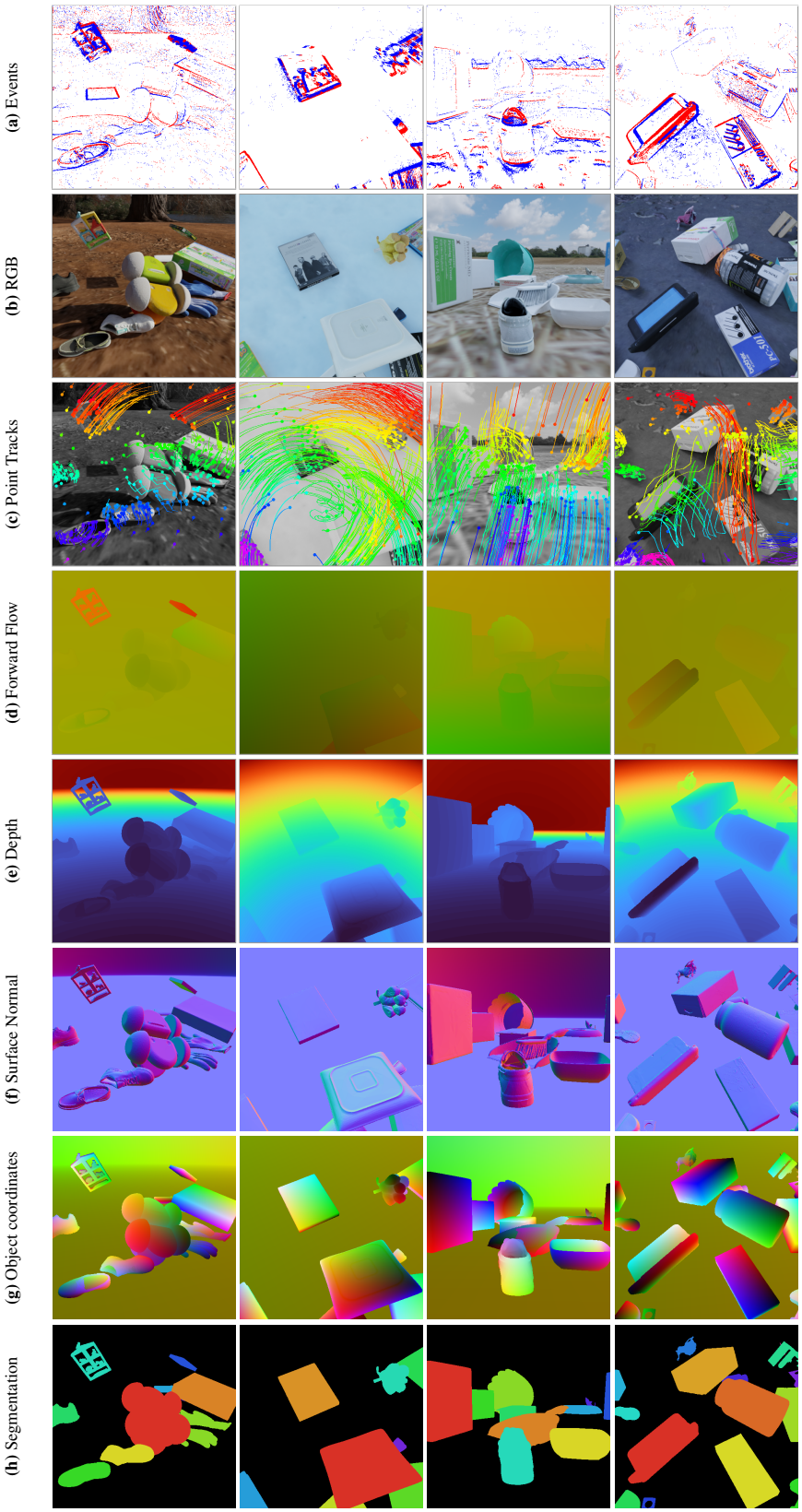

Figure 13. A few examples of EventKubric. Point tracks are subsampled for better visualization.

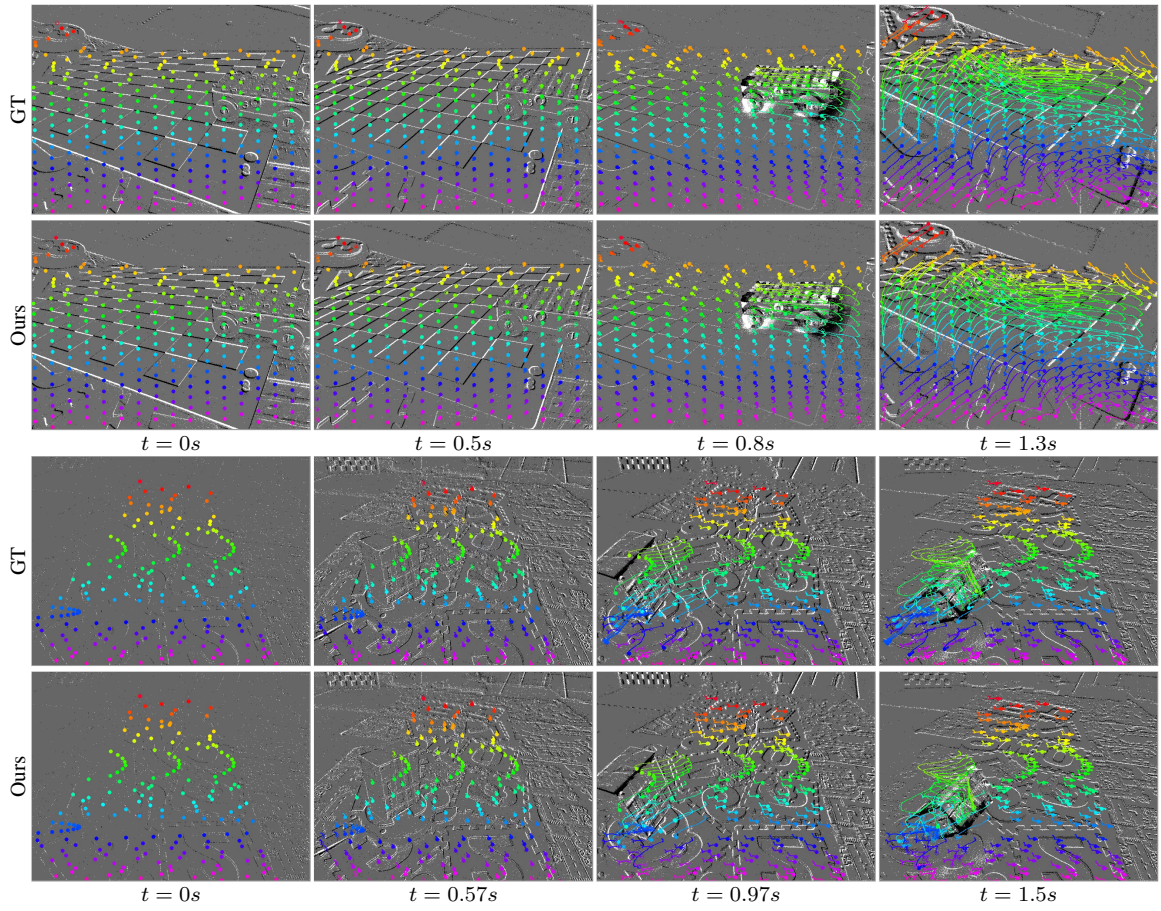

Figure 14. *Task 1 - TAP on EVIMO2 data*. Visualization of track predictions.

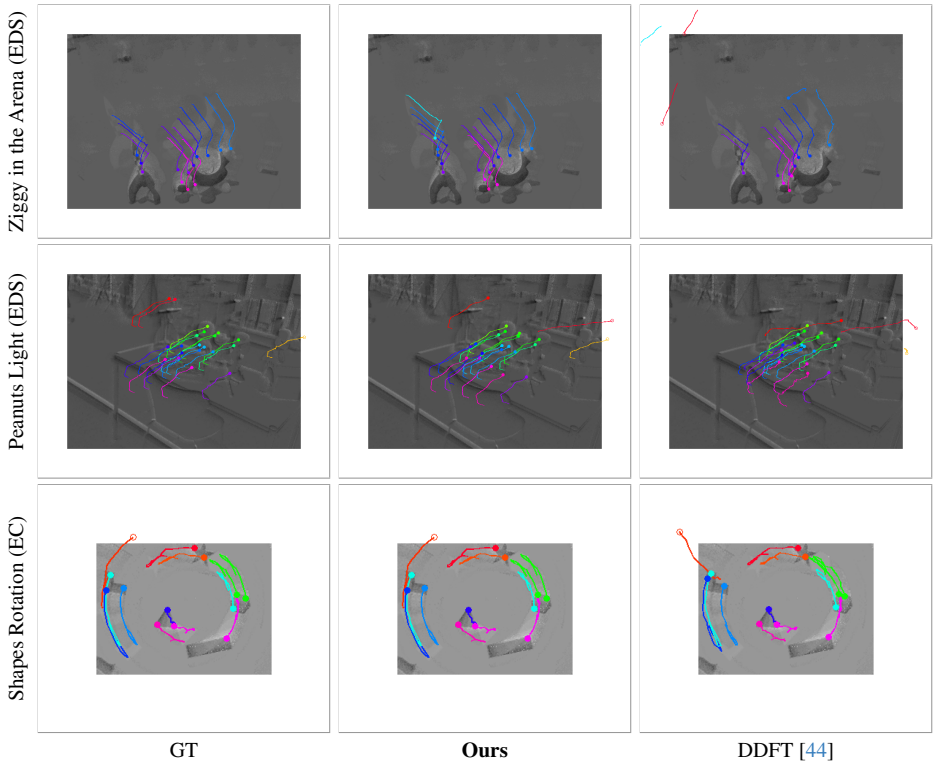

Figure 15. Additional visualizations on the EDS and EC dataset.

References

[1] Ignacio Alzugaray. *Event-driven Feature Detection and Tracking for Visual SLAM*. PhD thesis, ETH Zurich, 2022. 2

[2] Ignacio Alzugaray and Margarita Chli. ACE: An efficient asynchronous corner tracker for event cameras. In *Int. Conf. 3D Vision (3DV)*, pages 653–661, 2018. 3

[3] Ignacio Alzugaray and Margarita Chli. Haste: multi-hypothesis asynchronous speeded-up tracking of events. In *British Mach. Vis. Conf. (BMVC)*, page 744, 2020. 7, 11

[4] Simon Baker, Ralph Gross, Ishikawa Takahiro, and Iain Matthews. Lucas-kanade 20 years on: A unifying framework: Part 2. *Technical Report CMU-RI-TR-03-01*, 2003. 2

[5] D Blender Online Community. Blender—a 3d modelling and rendering package. *Blender Foundation*, 2018. 5

[6] Levi Burner, Anton Mitrokhin, Cornelia Fermüller, and Yiannis Aloimonos. EVIMO2: An event camera dataset for motion segmentation, optical flow, structure from motion, and visual inertial odometry in indoor scenes with monocular or stereo algorithms. *arXiv e-prints*, 2022. 6

[7] Weirong Chen, Le Chen, Rui Wang, and Marc Pollefeys. LEAP-VO: Long-term effective any point tracking for visual odometry. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 19844–19853, 2024. 1

[8] Feng Cheng and Gedas Bertasius. TallFormer: Temporal action localization with a long-memory transformer. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 503–521, 2022. 3

[9] Philippe Chiberre, Etienne Perot, Amos Sironi, and Vincent Lepetit. Detecting stable keypoints from events through image gradient prediction. In *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, 2021. 3

[10] Seokju Cho, Jiahui Huang, Jisu Nam, Honggyu An, Seungryong Kim, and Joon-Young Lee. Local all-pair correspondence for point tracking. *Eur. Conf. Comput. Vis. (ECCV)*, 2024. 3

[11] Erwin Coumans. Bullet physics simulation. In *ACM SIGGRAPH 2015 Courses*, 2015. 5

[12] Yongjian Deng, Hao Chen, Hai Liu, and Youfu Li. A voxel graph cnn for object classification with event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1172–1181, 2022. 4

[13] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. TAP-Vid: A benchmark for tracking any point in a video. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, pages 13610–13626, 2022. 1, 3, 6

[14] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Int. Conf. Comput. Vis. (ICCV)*, pages 10061–10072, 2023. 3, 6

[15] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Int. Conf. Comput. Vis. (ICCV)*, pages 2758–2766, 2015. 1, 3

[16] Tobias Fischer, Jiangmiao Pang, Thomas E Huang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *arXiv preprint arXiv:2210.06984*, 2022. 3

[17] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):154–180, 2022. 3, 9

[18] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Asynchronous, photometric feature tracking using events and frames. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 766–781, 2018. 3

[19] Daniel Gehrig, Antonio Loquercio, Konstantinos G. Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Int. Conf. Comput. Vis. (ICCV)*, pages 5632–5642, 2019. 4

[20] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to Events: Recycling video datasets for event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3583–3592, 2020. 2, 6

[21] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. EKLT: Asynchronous photometric feature tracking using events and frames. *Int. J. Comput. Vis.*, 128: 601–618, 2020. 3, 7, 11

[22] Mathias Gehrig, Manasi Muglikar, and Davide Scaramuzza. Dense continuous-time optical flow from event cameras. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(7):4736–4746, 2024. 1, 3, 4, 8

[23] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3749–3761, 2022. 2, 3, 5

[24] Friedhelm Hamann, Ziyun Wang, Ioannis Asmanis, Kenneth Chaney, Guillermo Gallego, and Kostas Daniilidis. Motion-prior contrast maximization for dense continuous-time motion estimation. In *Eur. Conf. Comput. Vis. (ECCV)*, 2024. 3, 6

[25] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 59–75, 2022. 1, 3

[26] Javier Hidalgo-Carrió, Guillermo Gallego, and Davide Scaramuzza. Event-aided direct sparse odometry. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5781–5790, 2022. 2, 7

[27] Berthold K.P. Horn and Brian G. Schunck. Determining optical flow. *J. Artificial Intell.*, 17(1):185 – 203, 1981. 3

[28] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1647–1655, 2017. 3

[29] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is better to track together. *Eur. Conf. Comput. Vis. (ECCV)*, 2024. 3, 4, 5, 6, 7

[30] Simon Klenk, Marvin Motzet, Lukas Koestler, and Daniel Cremers. Deep event visual odometry. In *Int. Conf. 3D Vision (3DV)*, pages 739–749, 2024. 6, 8

[31] Matej Kristan, Jiří Matas, Aleš Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Cehovin, Alan Lukežič, et al. The ninth visual object tracking vot2021 challenge results. In *Int. Conf. Comput. Vis. (ICCV)*, pages 2711–2738, 2021. 3

[32] Beat Kueng, Elias Mueggler, Guillermo Gallego, and Davide Scaramuzza. Low-latency visual odometry using event-based feature tracks. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, pages 16–23, 2016. 3, 7, 11

[33] Xavier Lagorce, Cédric Meyer, Sio-Hoi Ieng, David Filliat, and Ryad Benosman. Asynchronous event-based multikernel algorithm for high-speed visual features tracking. *IEEE Trans. Neural Netw. Learn. Syst.*, 26(8):1710–1720, 2015. 3

[34] Xi Li, Weiming Hu, Chunhua Shen, Zhongfei Zhang, Anthony Dick, and Anton Van Den Hengel. A survey of appearance models in visual object tracking. *ACM transactions on Intelligent Systems and Technology (TIST)*, 4(4):1–48, 2013. 2

[35] Yijin Li, Zhaoyang Huang, Shuo Chen, Xiaoyu Shi, Hongsheng Li, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. Blinkflow: A dataset to push the limits of event-based optical flow estimation. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, pages 3881–3888, 2023. 3

[36] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE J. Solid-State Circuits*, 43(2):566–576, 2008. 3

[37] Martin Litzenberger, Christoph Posch, D. Bauer, Ahmed Nabil Belbachir, P. Schön, B. Kohn, and H. Garn. Embedded vision system for real-time object tracking using an asynchronous transient vision sensor. In *Digital Signal Processing Workshop*, pages 173–178, 2006. 3

[38] Jiaxiong Liu, Bo Wang, Zhen Tan, Jinpu Zhang, Hui Shen, and Dewen Hu. Tracking any point with frame-event fusion network at high frame rate. *arXiv preprint arXiv:2409.11953*, 2024. 3, 7, 11

[39] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[40] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Int. Joint Conf. Artificial Intell. (IJCAI)*, pages 674–679, 1981. 2

[41] 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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019. 3

[42] Iain Matthews, Takahiro Ishikawa, and Simon Baker. The template update problem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2004. 2

[43] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4040–4048, 2016. 3

[44] Nico Messikommer, Carter Fang, Mathias Gehrig, and Davide Scaramuzza. Data-driven feature tracking for event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 1, 3, 7, 11, 14

[45] 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. *Int. J. Robot. Research*, 36(2):142–149, 2017. 7

[46] Yeongwoo Nam, Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Stereo depth from events cameras: Concentrate and focus on the future. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 6104–6113, 2022. 4, 9

[47] Zhenjiang Ni, Sio-Hoï Ieng, Christoph Posch, Stéphane Régnier, and Ryad Benosman. Visual tracking using neuromorphic asynchronous event-based cameras. *Neural Computation*, 27(4):925–953, 2015. 3

[48] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021. 3

[49] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 3

[50] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. An asynchronous time-based image sensor. In *IEEE Int. Symp. Circuits Syst. (ISCAS)*, pages 2130–2133, 2008. 3

[51] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an open event camera simulator. In *Conf. on Robotics Learning (CoRL)*, pages 969–982. PMLR, 2018. 5, 6

[52] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019. 5, 6

[53] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. FILM: Frame interpolation for large motion. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 250–266, 2022. 5

[54] Peter Sand and Seth Teller. Particle video: Long-range motion estimation using point trajectories. *Int. J. Comput. Vis.*, 80:72–91, 2008. 1, 3

[55] Shintaro Shiba, Friedhelm Hamann, Yoshimitsu Aoki, and Guillermo Gallego. Event-based background oriented schlieren. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(4): 2011–2026, 2024. 3

[56] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 534–549, 2020. 8

[57] David Tedaldi, Guillermo Gallego, Elias Mueggler, and Davide Scaramuzza. Feature detection and tracking with the dynamic and active-pixel vision sensor (DAVIS). In *Int. Conf.*

Event-Based Control, Comm. Signal Proc. (EBCCSP), 2016. 3

[58] Zachary Teed and Jia Deng. RAFT: Recurrent all pairs field transforms for optical flow. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 402–419, 2020. 1, 3

[59] Mel Vecerik, Carl Doersch, Yi Yang, Todor Davchev, Yusuf Aytar, Guangyao Zhou, Raia Hadsell, Lourdes Agapito, and Jon Scholz. Robotap: Tracking arbitrary points for few-shot visual imitation. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 5397–5403, 2024. 1

[60] Ziyun Wang, Friedhelm Hamann, Kenneth Chaney, Wen Jiang, Guillermo Gallego, and Kostas Daniilidis. Event-based continuous color video decompression from single frames. *arXiv preprint arXiv:2312.00113*, 2023. 2, 6

[61] Qiangqiang Wu, Tianyu Yang, Wei Wu, and Antoni B Chan. Scalable video object segmentation with simplified framework. In *Int. Conf. Comput. Vis. (ICCV)*, pages 13879–13889, 2023. 3

[62] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Int. Conf. Comput. Vis. (ICCV)*, pages 19855–19865, 2023. 3

[63] Alex Zihao Zhu, Nikolay Atanasov, and Kostas Daniilidis. Event-based feature tracking with probabilistic data association. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 4465–4470, 2017. 3, 7, 11

[64] Nikola Zubic, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. From Chaos Comes Order: Ordering Event Representations for Object Recognition and Detection . In *Int. Conf. Comput. Vis. (ICCV)*, pages 12800–12810, 2023. 4