Motion-prior Contrast Maximization for Dense Continuous-Time Motion Estimation

Friedhelm Hamann[1], Ziyun Wang[2], Ioannis Asmanis[2], Kenneth Chaney[2], Guillermo Gallego[1,3], and Kostas Daniilidis[2,4]

[1] TU Berlin and SCIoI Excellence Cluster, Berlin, Germany
[2] University of Pennsylvania, Philadelphia, US
[3] Einstein Center Digital Future and Robotics Institute Germany, Berlin, Germany
[4] Archimedes, Athena RC, Greece

Supplementary

Project page

See https://github.com/tub-rip/NonlinearCMax.

Loss Details

Our proposed loss function furthermore has the following details, which we experimentally found to have a slight advantage over not using them.
Scaling warped events by time to t_{ref}. The contribution of each warped event to the IWE is weighted by its distance $\Delta t = |t_{\mathrm{ref}} - t_k|$.
Masking image border. We mask events transported outside the image plane by the warp. They do not contribute to the loss.
Polarity-split IWEs. We treat positive and negative events separately when calculating the IWE. Effectively, at each iteration, two IWEs are evaluated.

Results on MultiFlow dataset

Table 1 shows results on the MultiFlow dataset [6], which are coherent with the results on EVIMO. Our self-supervised method performs second best after models directly supervised on the ground truth trajectories. Similarly, Bézier curves have a slight performance advantage, over the other tested trajectories. A model trained with additional frames as input shows an improved performance over the event-only model showing potential for multi-domain extensions.

Runtime of Submodules

The first column of Tab. 4 in the main paper reports inference times of different methods. Here is a breakdown of module times for our method. network inference (Main Fig. 2c)): 7.27ms, computation of trajectories (Fig. 2c to 2d): 0.19ms. For the loss module (Main Fig. 2e), used only at training time: flow interpolation (Main Fig. 2e1 to 2e3): 86ms, event warping (Main Fig. 2e4): 7.18ms, building IWE (Main Fig. 2e5): 8.34ms Notably, the flow interpolation is the slowest step but is only required during training, not during inference.

Table 1: Results on MultiFlow dataset [6]. “Frames” indicates whether frames at t_s and t_e were used as additional input to the artificial neural network or not.

	Method	Frames	TEPE ↓	TAE ↓	%Out ↓
SL	BFlow, polyn.	✗	1.71	5.93	0.09
SL	BFlow, Bézier [6]	✗	1.68	5.87	0.09
SSL	Paredes et. al [10]	✗	14.81	61.14	0.84
SSL	Ours, polyn.	✗	8.57	31.38	0.48
SSL	Ours, Bézier	✗	8.15	29.89	0.46
SSL	Ours, Bézier	✓	7.27	27.76	0.43

Ablations on EVIMO2

Table 2 shows additional ablations on EVIMO2 (analogue to the tests on DSEC in Tab. 5 of the paper). The results confirm most design choices, like the number of neighbors, and the use of a randomized reference time. As the prediction time is longer in this dataset, we also see the influence of the motion prior type and degree. We found Bézier curves with degree $N_c = 10$ to work best.

Table 2: Sensitivity and ablation study on EVIMO2 data. Ours corresponds to “Ours, Bézier” in Tab. 3 of the main paper. Configurations marked with “–” are unchanged from our main result.

	N_{traj}	N_{tref}	N_c	Motion prior	TEPE ↓	TAE↓	%Out ↓
Ours	32	$\sim\mathcal{U}(0,1)$	10	Bézier	6.14	16.98	0.25
	8	–	–	–	6.63	18.11	0.26
	64	–	–	–	6.61	19.18	0.29
	–	1	–	–	6.76	18.32	0.26
	–	–	5	–	6.25	17.81	0.25
	–	–	30	–	6.44	17.95	0.28
	–	–	1	polynomial	7.97	21.98	0.39

Results on MVSEC

For completeness, Tab. 3 provides a qualitative comparison of our method with state-of-the-art techniques on MVSEC data [14]. The input for a sample consists of all event data between two consecutive frames (GT depth from the LiDAR is temporally upsampled to the frame rate of the DAVIS346 event cameras used, at 45Hz [12, 15]). The models were trained with the same hyperparameters as reported for the DSEC dataset. The results confirm the good performance of our model and outperforms most baseline methods. It performs on average 14% better than Paredes et al. [10] and 13% worse than the test-time optimization-based method from Shiba et al. [11].

Table 3: Quantitative evaluation on MVSEC data [14]. Best in bold, runner-up underlined. SL: supervised learning; SSL$_{\mathrm{F}}$: SSL trained with grayscale images; SSL$_{\mathrm{E}}$: SSL trained with events; MB: model-based methods.

		indoor_flying1		indoor_flying2		indoor_flying3	
		EPE↓	%$_{\mathrm{3PE}}$↓	EPE↓	%$_{\mathrm{3PE}}$↓	EPE↓	%$_{\mathrm{3PE}}$↓
SL	EV-FlowNet+ [13]	0.56	1.00	0.66	1.00	0.59	1.00
	E-RAFT [5]	-	-	-	-	-	-
	EV-FlowNet [5]	-	-	-	-	-	-
	TMA [8]	1.06	3.63	1.81	27.29	1.58	23.26
	Cuadrado *et al.* [3]	0.58	-	0.72	-	0.67	-
SSL$_{\mathrm{F}}$	EV-FlowNet [15]	1.03	2.20	1.72	15.1	1.53	11.9
	Ziluo *et al.* [4]	0.57	0.10	0.79	1.60	0.72	1.30
SSL$_{\mathrm{E}}$	EV-FlowNet [16]	0.58	**0.00**	1.02	4.00	0.87	3.00
	EV-FlowNet [9]	0.79	1.20	1.40	10.9	1.18	7.40
	EV-FlowNet [11]	-	-	-	-	-	-
	ConvGRU-EV-FlowNet [7]	0.60	0.51	1.17	8.06	0.93	5.64
	Paredes *et al.* [10]	0.44	**0.00**	0.88	4.51	0.70	2.41
	Ours	0.45	0.09	0.71	2.40	0.6	0.93
MB	Akolkar *et al.* [1]	1.52	-	1.59	-	1.89	-
	Brebion *et al.* [2]	0.52	0.10	0.98	5.50	0.71	2.10
	Shiba *et al.* [11]	**0.42**	0.09	**0.60**	**0.59**	**0.50**	**0.29**

Additional Qualitative Results on EVIMO2 and DSEC

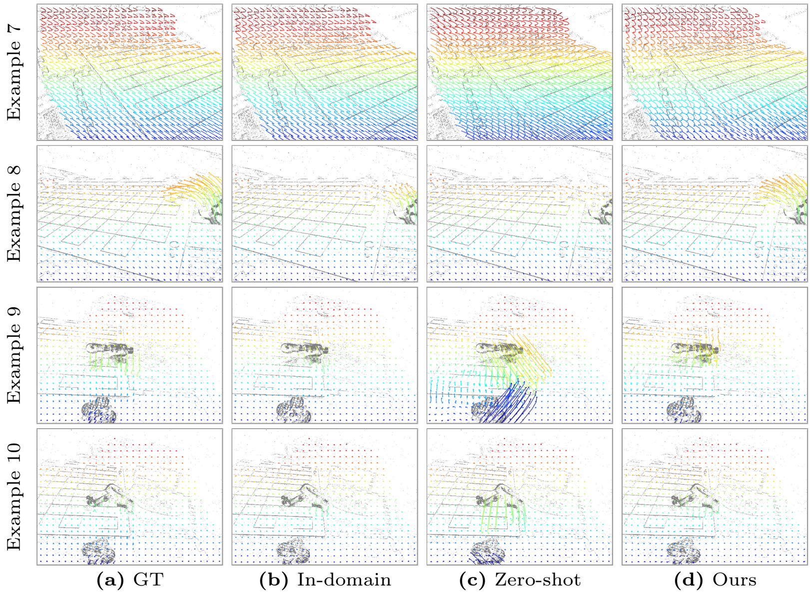

Fig. 1: Additional results on EVIMO2. Same notation as Fig. 3 in the main paper.

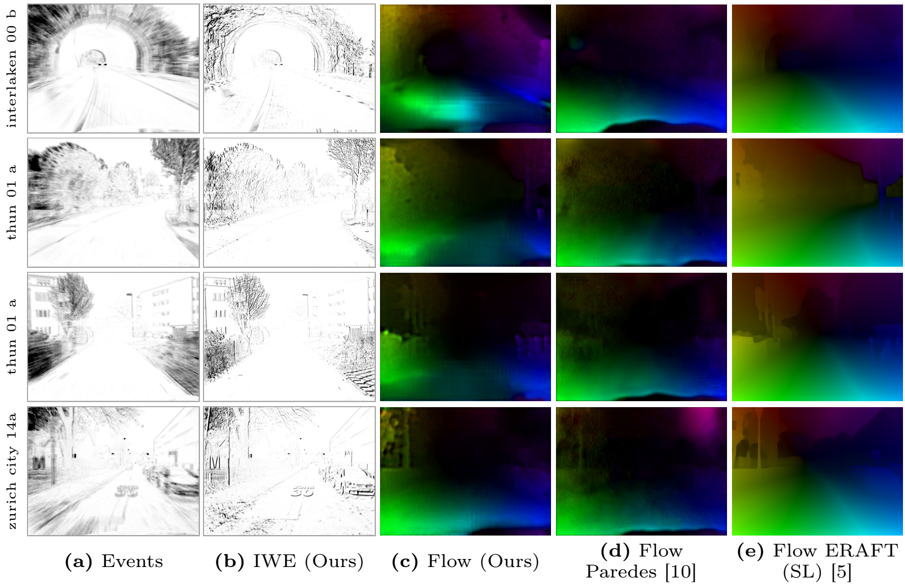

Fig. 2: Additional results on DSEC. Same notation as Fig. 5 in the main paper.

References

1. Akolkar, H., Ieng, S.H., Benosman, R.: Real-time high speed motion prediction using fast aperture-robust event-driven visual flow. IEEE Trans. Pattern Anal. Mach. Intell. **44**(1), 361–372 (2022). https://doi.org/10.1109/TPAMI.2020.3010468
2. Brebion, V., Moreau, J., Davoine, F.: Real-time optical flow for vehicular perception with low- and high-resolution event cameras. IEEE Trans. Intell. Transport. Syst. pp. 1–13 (2021). https://doi.org/10.1109/TITS.2021.3136358
3. Cuadrado, J., Rançon, U., Cottereau, B.R., Barranco, F., Masquelier, T.: Optical flow estimation from event-based cameras and spiking neural networks. Front. Neurosci. **17**, 1160034 (2023). https://doi.org/10.3389/fnins.2023.1160034
4. Ding, Z., Zhao, R., Zhang, J., Gao, T., Xiong, R., Yu, Z., Huang, T.: Spatio-temporal recurrent networks for event-based optical flow estimation. In: AAAI Conf. Artificial Intell. vol. 36, pp. 525–533 (2022). https://doi.org/10.1609/aaai.v36i1.19931
5. Gehrig, M., Millhäusler, M., Gehrig, D., Scaramuzza, D.: E-RAFT: Dense optical flow from event cameras. In: Int. Conf. 3D Vision (3DV). pp. 197–206 (2021). https://doi.org/10.1109/3DV53792.2021.00030
6. Gehrig, M., Muglikar, M., Scaramuzza, D.: Dense continuous-time optical flow from event cameras. IEEE Trans. Pattern Anal. Mach. Intell. pp. 1–12 (2024). https://doi.org/10.1109/TPAMI.2024.3361671
7. Hagenaars, J., Paredes-Vallés, F., De Croon, G.: Self-supervised learning of event-based optical flow with spiking neural networks. Adv. Neural Inf. Process. Syst. (NeurIPS) **34**, 7167–7179 (2021)
8. Liu, H., Chen, G., Qu, S., Zhang, Y., Li, Z., Knoll, A., Jiang, C.: TMA: Temporal motion aggregation for event-based optical flow. In: Int. Conf. Comput. Vis. (ICCV). pp. 9651–9660 (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.00888
9. Paredes-Valles, F., de Croon, G.C.H.E.: Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 3445–3454 (2021). https://doi.org/10.1109/CVPR46437.2021.00345
10. Paredes-Vallés, F., Scheper, K.Y., De Wagter, C., de Croon, G.C.: Taming contrast maximization for learning sequential, low-latency, event-based optical flow. In: Int. Conf. Comput. Vis. (ICCV). pp. 9661–9671 (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.00889
11. Shiba, S., Aoki, Y., Gallego, G.: Secrets of event-based optical flow. In: Eur. Conf. Comput. Vis. (ECCV). pp. 628–645 (2022). https://doi.org/10.1007/978-3-031-19797-0_36
12. Shiba, S., Klose, Y., Aoki, Y., Gallego, G.: Secrets of event-based optical flow, depth, and ego-motion by contrast maximization. IEEE Trans. Pattern Anal. Mach. Intell. pp. 1–18 (2024). https://doi.org/10.1109/TPAMI.2024.3396116
13. Stoffregen, T., Scheerlinck, C., Scaramuzza, D., Drummond, T., Barnes, N., Kleeman, L., Mahony, R.: Reducing the sim-to-real gap for event cameras. In: Eur. Conf. Comput. Vis. (ECCV). pp. 534–549 (2020). https://doi.org/https://doi.org/10.1007/978-3-030-58583-9_32
14. Zhu, A.Z., Thakur, D., Ozaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K.: The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. IEEE Robot. Autom. Lett. **3**(3), 2032–2039 (Jul 2018). https://doi.org/10.1109/lra.2018.2800793

15. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: EV-FlowNet: Self-supervised optical flow estimation for event-based cameras. In: Robotics: Science and Systems (RSS). pp. 1–9 (2018). https://doi.org/10.15607/RSS.2018.XIV.062
16. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 989–997 (2019). https://doi.org/10.1109/CVPR.2019.00108