Supplementary Material – Back to Event Basics: Self-Supervised Learning of Image Reconstruction for Event Cameras via Photometric Constancy

Federico Paredes-Vallés Guido C. H. E. de Croon
Micro Air Vehicle Laboratory, Delft University of Technology, The Netherlands

A. Sequence Cuts

The DAVIS frames accompanying the frequently used Event-Camera Dataset [1] usually suffer from motion blur and under/overexposure. For this reason, we only evaluate reconstruction accuracy on sections of this dataset in which the frames appear to be of high quality. The exact cut times are adopted from [2] and shown in Table 1. Additionally, we only evaluate optical flow accuracy on these sections to remain comparable to the results reported in [2].

Table 1: Sequence cuts used for evaluation on the Event-Camera Dataset [1]. Adopted from [2].

Sequence	Start [s]	End [s]
boxes_6dof_cut	5.0	20.0
calibration_cut	5.0	20.0
dynamic_6dof_cut	5.0	20.0
office_zigzag_cut	5.0	12.0
poster_6dof_cut	5.0	20.0
shapes_6dof_cut	5.0	20.0
slider_depth_cut	1.0	2.5

B. Impact of Event Deblurring

As discussed in this work, our self-supervised image reconstruction framework is designed around the event-based photometric constancy equation. While the right-hand side of this equation is obtained via the dot product between the warped spatial gradients of the last reconstructed image and the estimated optical flow; we propose that the left-hand side is obtained by integrating the deblurred (and averaged) input events. Since the main supervisory signal used to train our image reconstruction architectures comes from the comparison of the two sides of this equation, after training, the spatial gradients of the reconstructed images are correlated with the integrated events. These events, if not warped to the timestamp of the reconstructed frame, would introduce motion blur into the images. The amount of motion blur would depend on the density of events and on the length of the partition of events.

Figure 1: Qualitative evaluation of the impact of event deblurring on the quality of the reconstructed frames on sequences from the ECD [1] dataset.

Table 2: Quantitative evaluation of the impact of event deblurring prior to event integration on the ECD [1] and HQF [2] datasets. For each dataset, we report the mean MSE (↓), SSIM [3] (↑) and LPIPS [4] (↓). Best in bold.

	ECD*			HQF		
	MSE	SSIM	LPIPS	MSE	SSIM	LPIPS
E2VID$_E$ (w/ deblurring)	**0.06**	**0.55**	**0.37**	**0.06**	**0.48**	**0.47**
E2VID$_E$ (w/o deblurring)	0.14	0.30	0.58	0.11	0.28	0.64

*Sequence cuts in Table 1.

To validate this approach, we conducted an ablation study in which we trained the same ReconNet architecture (accompanied by the same pre-trained optical flow network) with and without event deblurring prior to event integration. Quantitative results are presented in Table 2, and are supported by qualitative results in Fig. 1. As shown, event deblurring is a crucial mechanism to reconstruct sharp images from the events. Without it, the reconstructed frames appear less sharp for the same number of input events, and the network is characterized by significantly worse error metrics on the evaluation datasets.

C. Additional Quantitative Results

A breakdown of the quantitative results of our FlowNet and ReconNet architectures on the ECD [1] and HQF [2] datasets can be found in Tables 3 and 4, respectively.

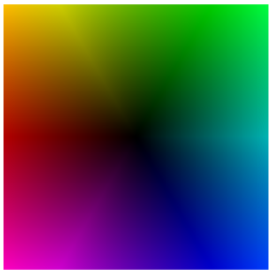


Figure 2: Optical flow field color-coding scheme. Direction is encoded incolor hue, and speed in color brightness.

D. Additional Qualitative Results

Figs. 3, 4, and 5 show additional qualitative results of our FlowNet and ReconNet architectures on the ECD [1] and HQF [2] datasets. Lastly, Fig. 6 shows qualitative results on the high-resolution automotive dataset recently released by Prophesee [5]. The optical flow color-coding scheme for Figs. 3 and 6 can be found in Fig. 2.

References

[1] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, and D. Scaramuzza, "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam," *Int. J. Robot. Research*, vol. 36, no. 2, pp. 142–149, 2017.

[2] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, T. Drummond, N. Barnes, L. Kleeman, and R. Mahony, "Reducing the sim-to-real gap for event cameras," in *European Conf. Comput. Vis. (ECCV)*, 2020.

[3] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[4] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conf. on Comput. Vis. Pattern Recog. (CVPR)*, 2018, pp. 586–595.

[5] 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 Process. Systems (NeurIPS)*, 2020.

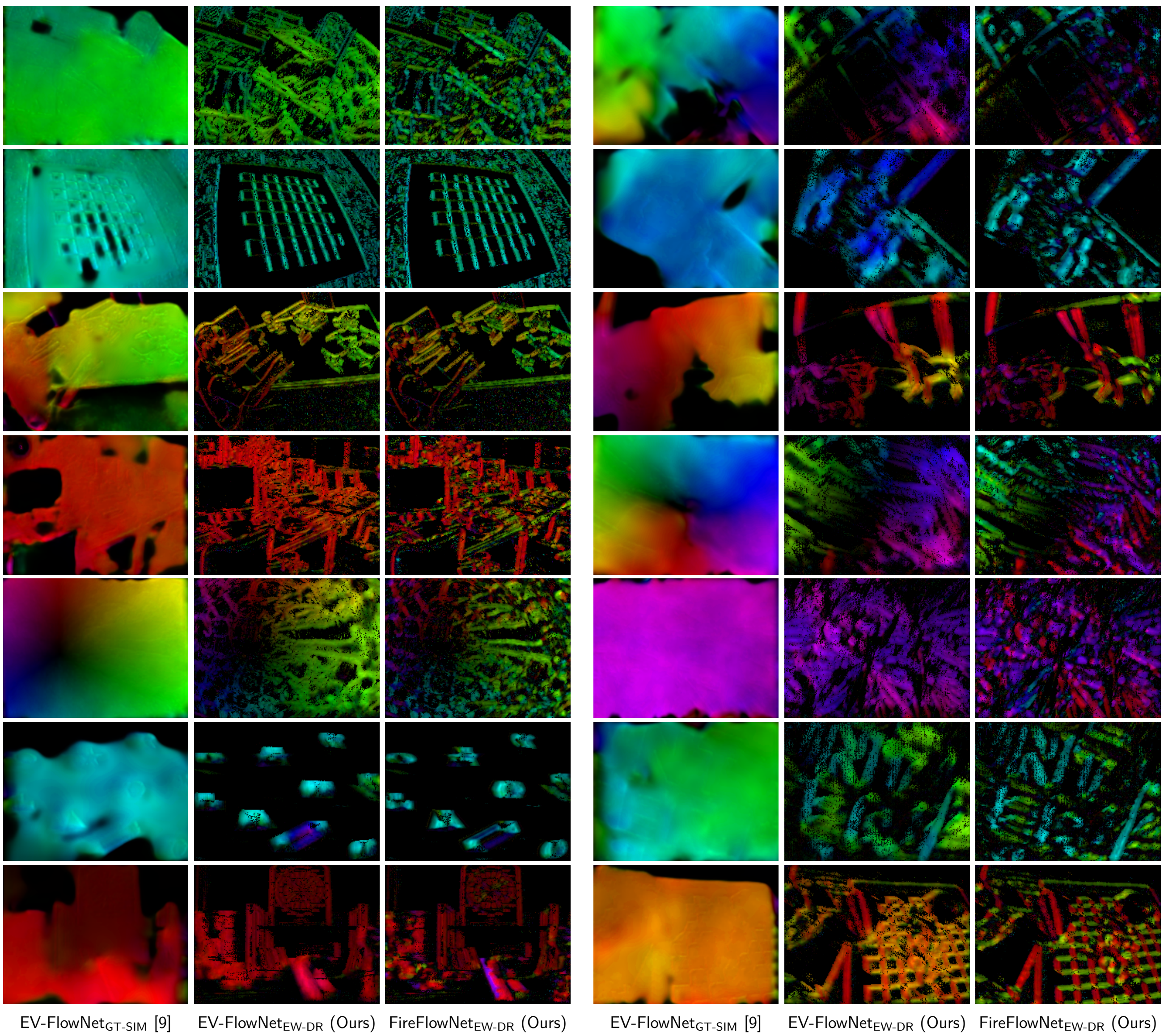
[6] A. Z. Zhu and L. Yuan, "EV-FlowNet: Self-supervised optical flow estimation for event-based cameras," in *Robot.: Science and Systems (RSS)*, 2018.

Table 3: Breakdown of the quantitative evaluation of our FlowNet architectures on the ECD [1] and HQF [2] datasets. For each dataset, we report the FWL [2] (↑).

	EV-FlowNet$_{\text{FW-MVSEC}}$ [6]	EV-FlowNet$_{\text{GT-SIM}}$ [2]	EV-FlowNet$_{\text{EW-DR}}$ (Ours)	FireFlowNet$_{\text{EW-DR}}$ (Ours)
		ECD*		
boxes_6dof_cut	1.42	1.46	1.22	1.37
calibration_cut	1.20	1.31	1.11	1.22
dynamic_6dof_cut	1.37	1.39	1.22	1.33
office_zigzag_cut	1.13	1.11	1.09	1.18
poster_6dof_cut	1.50	1.56	1.20	1.34
shapes_6dof_cut	1.15	1.57	1.51	1.38
slider_depth_cut	1.73	2.17	1.80	1.88
Mean	**1.36**	**1.51**	**1.31**	**1.39**
		HQF		
bike_day_hdr	1.22	1.23	1.49	1.52
boxes	1.75	1.80	1.68	1.72
desk	1.23	1.35	1.35	1.42
desk_fast	1.43	1.50	1.42	1.47
desk_hand_only	0.95	0.85	1.14	1.23
desk_slow	1.01	1.08	1.23	1.27
engineering_posters	1.50	1.65	1.65	1.71
high_texture_plants	0.13	1.68	1.71	1.77
poster_pillar_1	1.20	1.24	1.39	1.45
poster_pillar_2	1.16	0.96	1.10	1.18
reflective_materials	1.45	1.57	1.62	1.63
slow_and_fast_desk	0.93	0.99	1.68	1.77
slow_hand	1.64	1.56	1.90	1.96
still_life	1.93	1.98	1.76	1.97
Mean	**1.25**	**1.39**	**1.51**	**1.58**

Table 4: Breakdown of the quantitative results of our ReconNet architectures on the ECD [1] and HQF [2] datasets. For each sequence, we report the MSE (↓), SSIM [3] (↑) and LPIPS [4] (↓). The F and E subscripts determine whether our networks were trained in combination with FireFlowNet or EV-FlowNet, respectively.

	MSE				SSIM				LPIPS			
	FireNet_F	FireNet_E	E2VID_F	E2VID_E	FireNet_F	FireNet_E	E2VID_F	E2VID_E	FireNet_F	FireNet_E	E2VID_F	E2VID_E
						ECD*						
boxes_6dof_cut	0.0533	0.0554	0.0540	0.0541	0.5705	0.5538	0.5785	0.5997	0.3736	0.4170	0.3776	0.3781
calibration_cut	0.0531	0.0620	0.0779	0.0677	0.5464	0.5356	0.5445	0.5594	0.2770	0.3046	0.2982	0.2937
dynamic_6dof_cut	0.0950	0.0780	0.1030	0.0845	0.4037	0.4036	0.4123	0.4519	0.4773	0.4969	0.4576	0.4424
office_zigzag_cut	0.0452	0.0427	0.0442	0.0617	0.5019	0.5033	0.4970	0.4807	0.3634	0.4122	0.3350	0.3485
poster_6dof_cut	0.0592	0.0567	0.0593	0.0521	0.5385	0.5211	0.5613	0.5823	0.4039	0.4396	0.3941	0.3909
shapes_6dof_cut	0.0500	0.0928	0.0608	0.0594	0.5719	0.5262	0.5673	0.6297	0.4303	0.4313	0.4532	0.3554
slider_depth_cut	0.0612	0.0613	0.0840	0.0660	0.5200	0.5265	0.4758	0.5174	0.3613	0.3834	0.3536	0.3728
Mean	**0.0595**	**0.0641**	**0.0690**	**0.0636**	**0.5218**	**0.5100**	**0.5195**	**0.5459**	**0.3838**	**0.4121**	**0.3813**	**0.3688**
						HQF						
bike_day_hdr	0.0629	0.0587	0.0552	0.0519	0.4317	0.4471	0.4574	0.4835	0.5248	0.5584	0.5028	0.5266
boxes	0.0596	0.0549	0.0694	0.0562	0.4885	0.4912	0.4853	0.5190	0.3994	0.4439	0.4108	0.4164
desk	0.0619	0.0649	0.0817	0.0697	0.4776	0.4779	0.4677	0.4972	0.3938	0.4373	0.4018	0.3914
desk_fast	0.0588	0.0624	0.0711	0.0637	0.4935	0.4882	0.5027	0.5238	0.4482	0.4999	0.4425	0.4515
desk_hand_only	0.0805	0.0910	0.0755	0.0594	0.5143	0.5106	0.5134	0.5545	0.5971	0.6202	0.5619	0.5438
desk_slow	0.0783	0.0894	0.0976	0.0759	0.5011	0.4341	0.2852	0.4998	0.5214	0.6029	0.6689	0.5253
engineering_posters	0.0570	0.0541	0.0783	0.0656	0.4690	0.4776	0.4456	0.4797	0.4250	0.4417	0.4345	0.4528
high_texture_plants	0.0579	0.0581	0.0687	0.0653	0.4689	0.4705	0.4081	0.4404	0.3618	0.4054	0.3895	0.3825
poster_pillar_1	0.0653	0.0623	0.0726	0.0641	0.3132	0.3121	0.3340	0.3455	0.5532	0.5720	0.5144	0.5455
poster_pillar_2	0.0638	0.0605	0.0644	0.0532	0.3569	0.3814	0.3881	0.4119	0.5968	0.6059	0.5643	0.5737
reflective_materials	0.0506	0.0517	0.0566	0.0528	0.4621	0.4705	0.4779	0.5032	0.4235	0.4655	0.4254	0.4493
slow_and_fast_desk	0.0701	0.0648	0.0620	0.0699	0.4503	0.4584	0.4805	0.4850	0.4565	0.4903	0.4200	0.4321
slow_hand	0.0824	0.0667	0.0736	0.0614	0.4123	0.4246	0.4380	0.4647	0.5480	0.5651	0.4694	0.4937
still_life	0.0429	0.0419	0.0486	0.0469	0.5434	0.5413	0.5376	0.5470	0.3924	0.4400	0.4187	0.4515
Mean	**0.0637**	**0.0629**	**0.0696**	**0.0611**	**0.4559**	**0.4561**	**0.4444**	**0.4825**	**0.4744**	**0.5106**	**0.4732**	**0.4740**



(a) ECD dataset.

(b) HQF dataset.

Figure 3: Additional qualitative comparison of our FlowNet architectures with the state-of-the-art EV-FlowNet [2] on sequences from the ECD [1] and HQF [2] dataset.

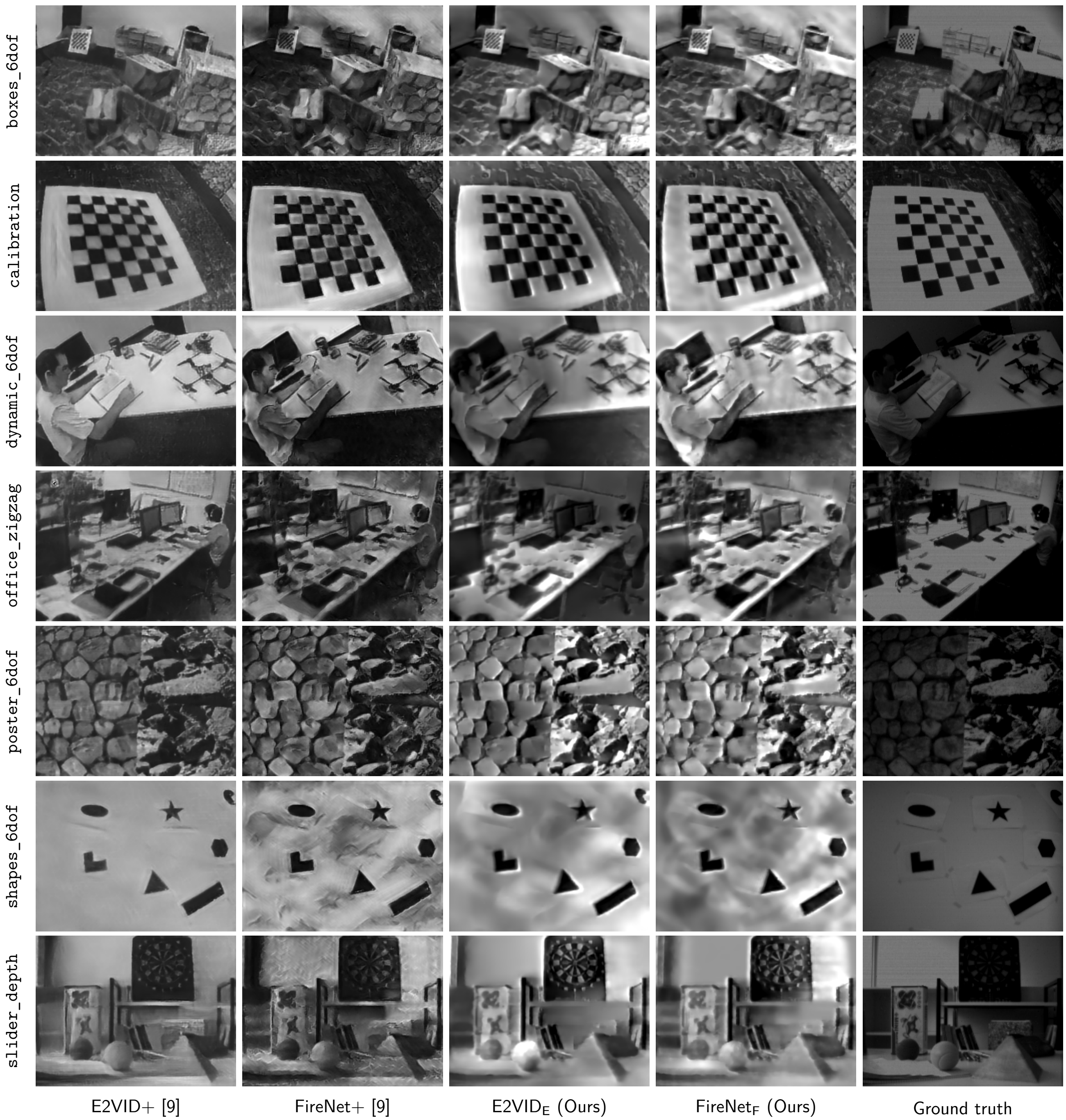


Figure 4: Additional qualitative comparison of our ReconNet architectures with the state-of-the-art E2VID+ and FireNet+ [2] on sequences from the ECD [1] dataset. Local histogram equalization not used for this comparison.


bike_bay_hdr
boxes
desk
desk_fast
desk_hand_only
desk_slow
engineering_posters
E2VID+ [9]
FireNet+ [9]
E2VID_E (Ours)
FireNet_F (Ours)
Ground truth

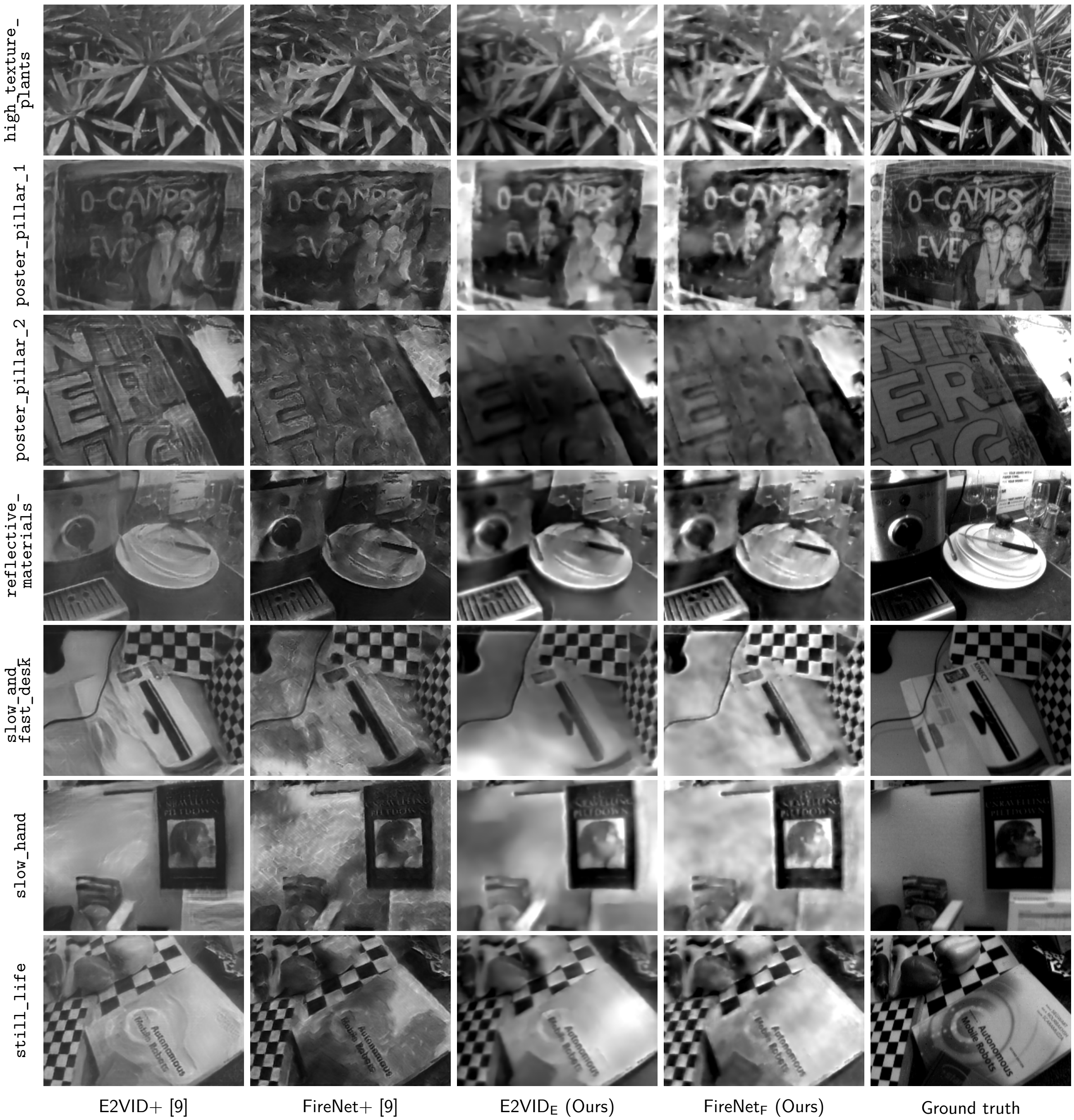
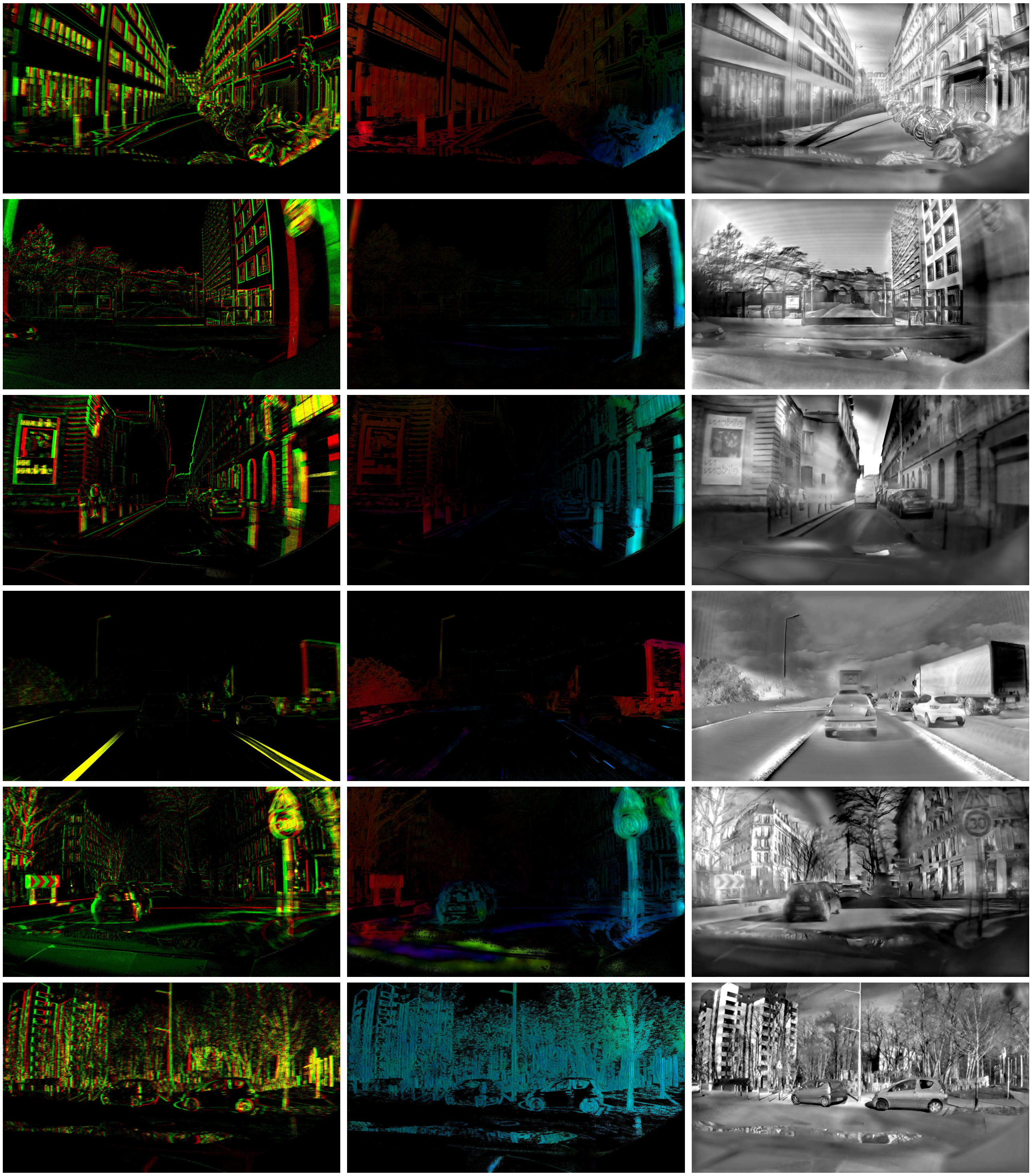


Figure 5: Additional qualitative comparison of our ReconNet architectures with the state-of-the-art E2VID+ and FireNet+ [2] on sequences from the HQF [2] dataset. Local histogram equalization not used for this comparison.



Input events   EV-FlowNet$_{\text{EW-DR}}$ E2VID$_{\text{E}}$

Figure 6: Additional qualitative results on sequences from Prophesee's high-resolution automotive dataset [5].