6 Supplementary Materials

6.1 HQF Dataset Details

The details of the High Quality Frames dataset are shown in Table 6.

Table 6: Breakdown of the sequences included in HQF. To provide some inter-device variability, the dataset is taken with two separate DAVIS 240C cameras, 1 and 2.

Sequence	Length [s]	Cam.	Frames [k]	Events [M]	Description
bike_bay_hdr	99.0	1	2.4	19.8	Camera moves from dim to bright
boxes	24.2	1	0.5	10.1	Indoor light, translations
desk	65.8	2	1.5	13.5	Natural light, various motions
desk_fast	32.0	2	0.7	12.6	Natural light, fast motions
desk_hand_only	20.6	2	0.5	0.8	Indoor light, static camera
desk_slow	63.3	2	1.4	1.9	Natural light, slow motions
engineering_posters	60.7	1	1.3	15.4	Indoor light, text and images
high_texture_plants	43.2	1	1.1	14.6	Outdoors, high textures
poster_pillar_1	41.8	1	1.0	7.1	Outdoors, text and images
poster_pillar_2	25.4	1	0.6	2.5	Outdoors, text and images, long pause
reflective_materials	28.9	1	0.6	7.8	Natural light, reflective objects
slow_and_fast_desk	75.6	1	1.7	15.0	Natural light, diverse motion
slow_hand	38.9	1	0.9	7.6	Indoor, slow motion, static camera
still_life	68.1	1	1.2	42.7	Indoors, Indoor light, 6DOF motions

6.2 Voxel Generation

Two natural choices for generating voxel grids from the event stream are *fixed rate* and *fixed events* (Figure 5). In fixed rate, voxels are formed from t second wide slices of the event stream (variable event count), endowing the resulting inference with a fixed frame rate. This has the downside that inference cannot adapt to changing scene dynamics, a disadvantage shared by conventional cameras. A special case of fixed rate is *between frames* where all events between two image frames are used to form a voxel grid.

In fixed events, one waits for N events before making a voxel grid no matter how long it takes (variable duration). Fixed events has the downside that if the camera receives few events, either because the scene has little texture or the motion is slow, the inference rate can slow to a crawl. This method allows matching the value N to the average N of the training set during inference, potentially benefiting the network. The average events per voxel in our training set is 0.0564.

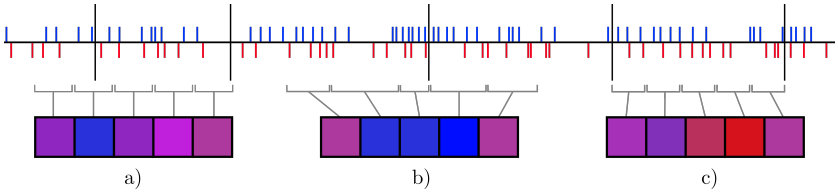


Fig. 5: Events (blue and red lines) on a timeline are discretized into voxels below (squares) according to: a) *fixed rate*, b) *fixed events*, c) *between frames* (frames denoted by black lines).

6.3 CTs extended

As outlined in Section 3, we propose several methods for estimating the CTs for a given event-based dataset. In total, we tried three different methods:

- Creating a simulator scene with similar texture and range of motions as the real sequence and adjusting the CTs until the $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ match.
- Simulating events from the APS frames (if they are available) and adjusting the CTs until the $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ match the real sequence.
- Creating a calibration scene in the simulator, recording this scene on-screen and adjusting the CT of the event camera to match the $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ of the simulation.

We describe the first two approaches in the main paper. These approaches indicate CTs of approximately 0.3 and 0.75 for IJRR and MVSEC respectively.

We also produced a calibration sequence in an attempt to match the simulator to our particular DAVIS 240C at default settings. For this, we moved a checkerboard across the image plane in ESIM, using various CTs. The scene was played on a high-refresh screen and recorded by our DAVIS. The resulting event-rate for each sequence, shown in Table 7, suggest a CT ≈ 0.5 to match the camera to the simulator. This is however in conflict with the $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ (8.2) of the real sequences compared to the $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ of the best training data CT (19.6). In other words, there seems to be a mismatch for this method of calibration, perhaps stemming from a difference in recording events from real scenes to recording scenes from a screen as was done for the checkerboard calibration sequence.

Table 7: Comparison of the $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ for simulated sequences at various CT settings with the $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ of a real calibration sequence. The sequence consists of a checkerboard in motion. The same sequence is also recorded by a real event camera (DAVIS 240C) using default bias settings. The result suggests that a CT value around 0.5 would be appropriate to match the simulator to the real camera.

Contrast threshold	0.2	0.5	0.75	1.0	1.5	Real
$\frac{\text{events}}{\text{pix}\cdot\text{s}}$	26.5	19.6	16.2	10.8	7.2	8.2
LPIPS	0.289	0.285	0.289	0.311	0.316	-

6.4 Sequence Cuts

Since the frames accompanying the events in the commonly used MVSEC and IJRR datasets are of low quality, we only evaluate on select sequences and for select cuts of those sequences. These cuts are enumerated in Table 8.

Table 8: Start and end times for sequences in IJRR and MVSEC that we present validation statistics on. While both IJRR and MVSEC contain more sequences than the ones listed, those not included had very low quality accompanying frames (see Figure 3).

IJRR			MVSEC		
Sequence	Start [s]	End [s]	Sequence	Start [s]	End [s]
boxes_6dof	5.0	20.0	indoor_flying1	10.0	70.0
calibration	5.0	20.0	indoor_flying2	10.0	70.0
dynamic_6dof	5.0	20.0	indoor_flying3	10.0	70.0
office_zigzag	5.0	12.0	indoor_flying4	10.0	19.8
poster_6dof	5.0	20.0	outdoor_day1	0.0	60.0
shapes_6dof	5.0	20.0	outdoor_day2	100.0	160.0
slider_depth	1.0	2.5			

6.5 MVSEC Expanded Results

For space reasons we did not include results of other works which quote MVSEC. This is because these works did not release the models, which meant we could not compare on the FWL metric, so we did not include them in the main paper. However, for completeness, we show the results of [39, 44] and [12]. As can be seen in Table 9, our network still compares very favorably. Interestingly, zero loss (doing nothing) is still the best overall at reducing outliers and is a strong

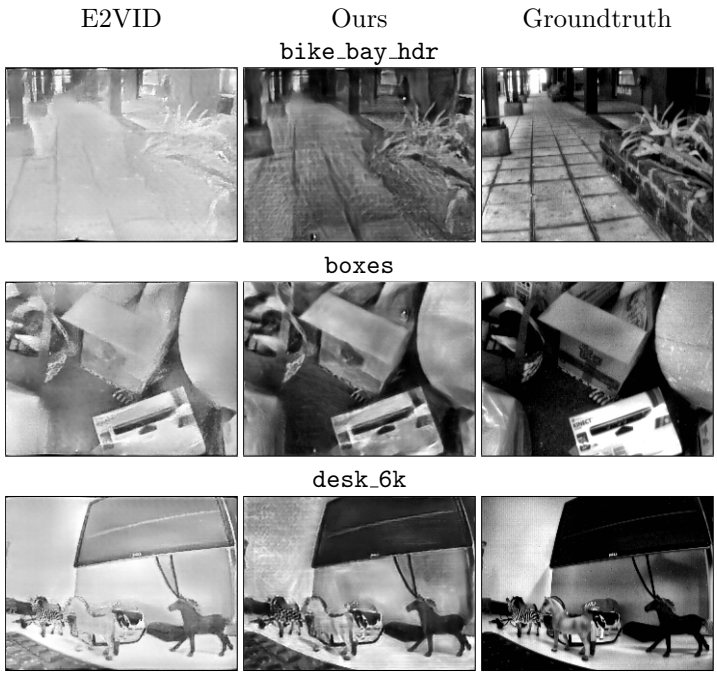
Table 9: Comparison of various methods to optic flow estimated from Lidar depth and ego-motion sensors [42]. The average-endpoint-error to the Lidar estimate (AEE) and the percentage of pixels with AEE above 3 and greater than 5 % of the magnitude of the flow vector ($\%_{\text{Outlier}}$) are presented for each method (lower is better, best in bold). The time between frames is dt=1. Zeros is the baseline error resulting from always estimating zero flow.

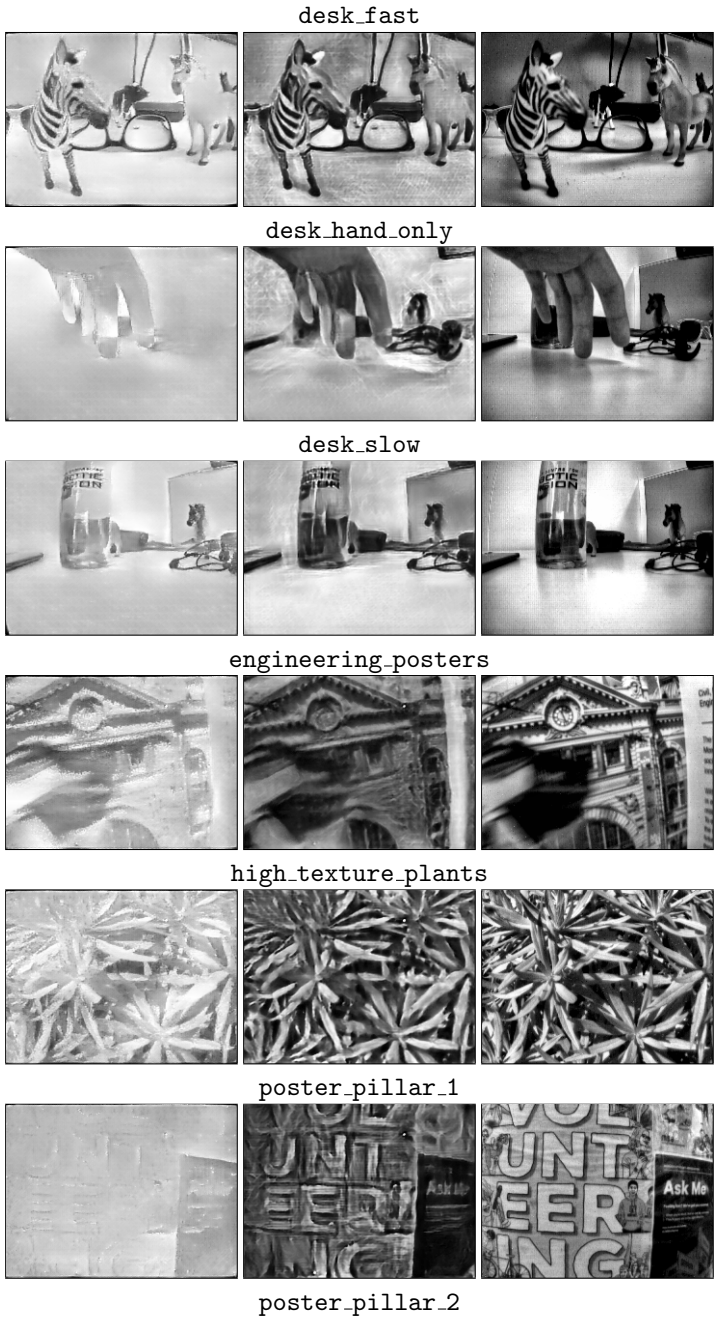
Dataset	outdoor_day1		outdoor_day2		indoor_flying1		indoor_flying2		indoor_flying3	
	AEE	$\%_{\text{Outlier}}$	AEE	$\%_{\text{Outlier}}$	AEE	$\%_{\text{Outlier}}$	AEE	$\%_{\text{Outlier}}$	AEE	$\%_{\text{Outlier}}$
Zeros	4.31	0.39	1.07	**0.91**	1.10	1.00	1.74	**0.89**	1.50	**0.94**
EVFlow [43]	0.49	0.20	-	-	1.03	2.20	1.72	15.10	1.53	11.90
EVFlow+ [44]	**0.32**	**0.00**	-	-	0.58	**0.00**	1.02	4.00	0.87	3.00
Gehrig [12]	-	-	-	-	0.96	0.91	1.38	8.20	1.40	6.47
Ours	0.68	0.99	**0.82**	0.96	**0.56**	1.00	**0.66**	1.00	**0.59**	1.00
ECN* [39]	0.35	0.04	-	-	0.21	0.01	-	-	-	-

*ECN is trained on 80 % of the sequence and evaluated on the remaining 20 %. This prevents direct comparison, however we include their result for completeness sake.

contender for AEE (especially in the flying sequences), showing the importance of reporting the relative improvement as in our FWL.

6.6 Additional Qualitative Results




desk_fast
desk_hand_only
desk_slow
engineering_posters
high_texture_plants
poster_pillar_1
poster_pillar_2

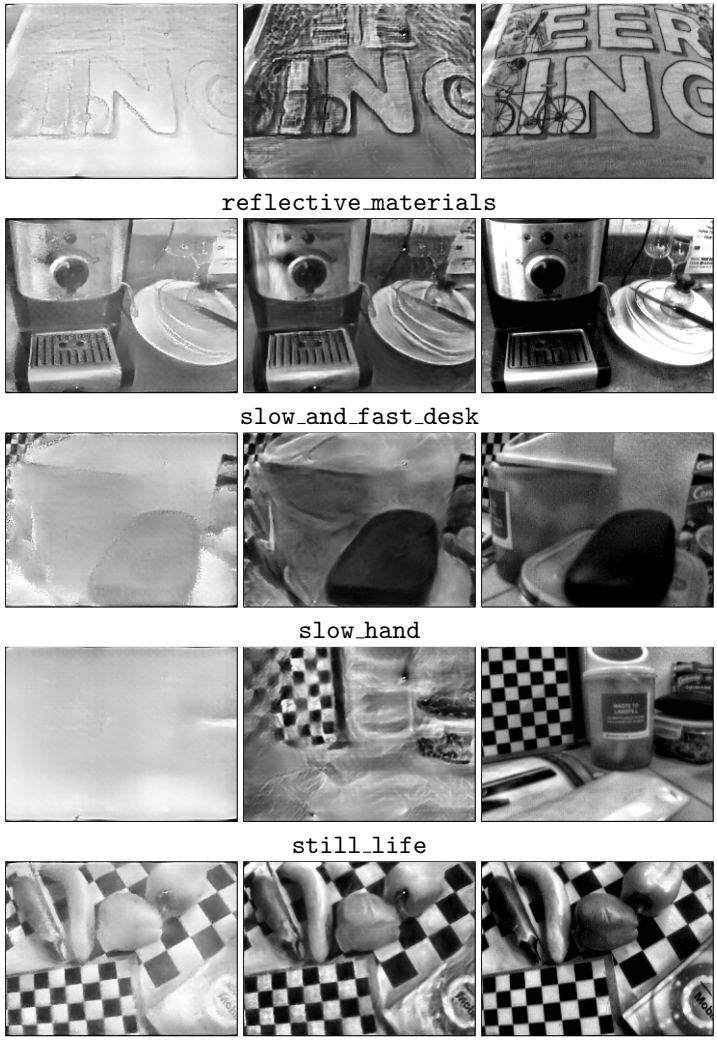
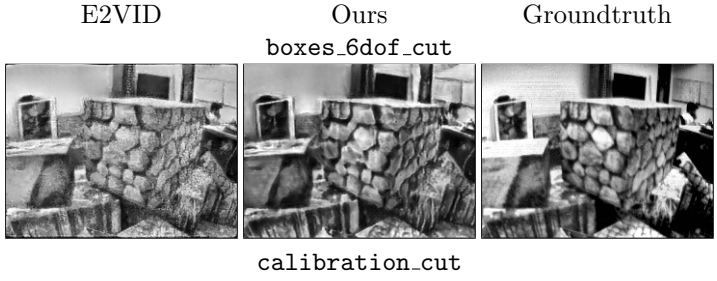


Table 10: Qualitative results for HQFD. Random selection, not cherry picked.



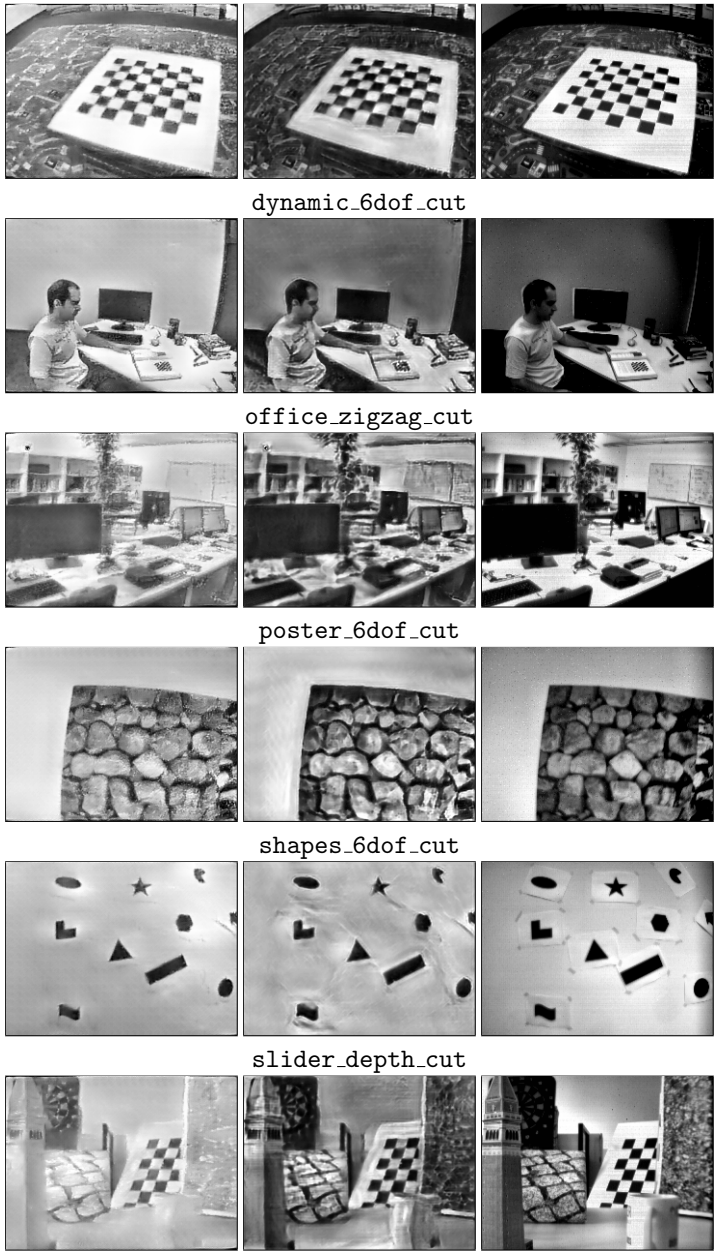


Table 11: Qualitative results for IJRR. Random selection, not cherry picked.

Table 12: Qualitative results for MVSEC. Random selection, not cherry picked.

E2VID Ours Groundtruth

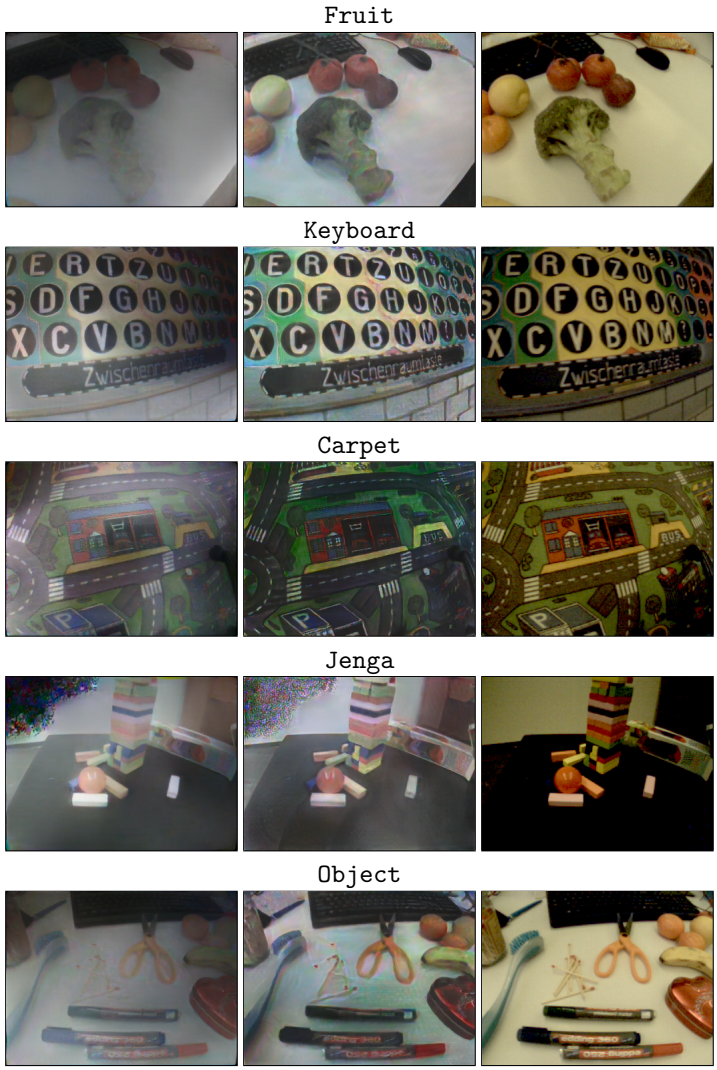
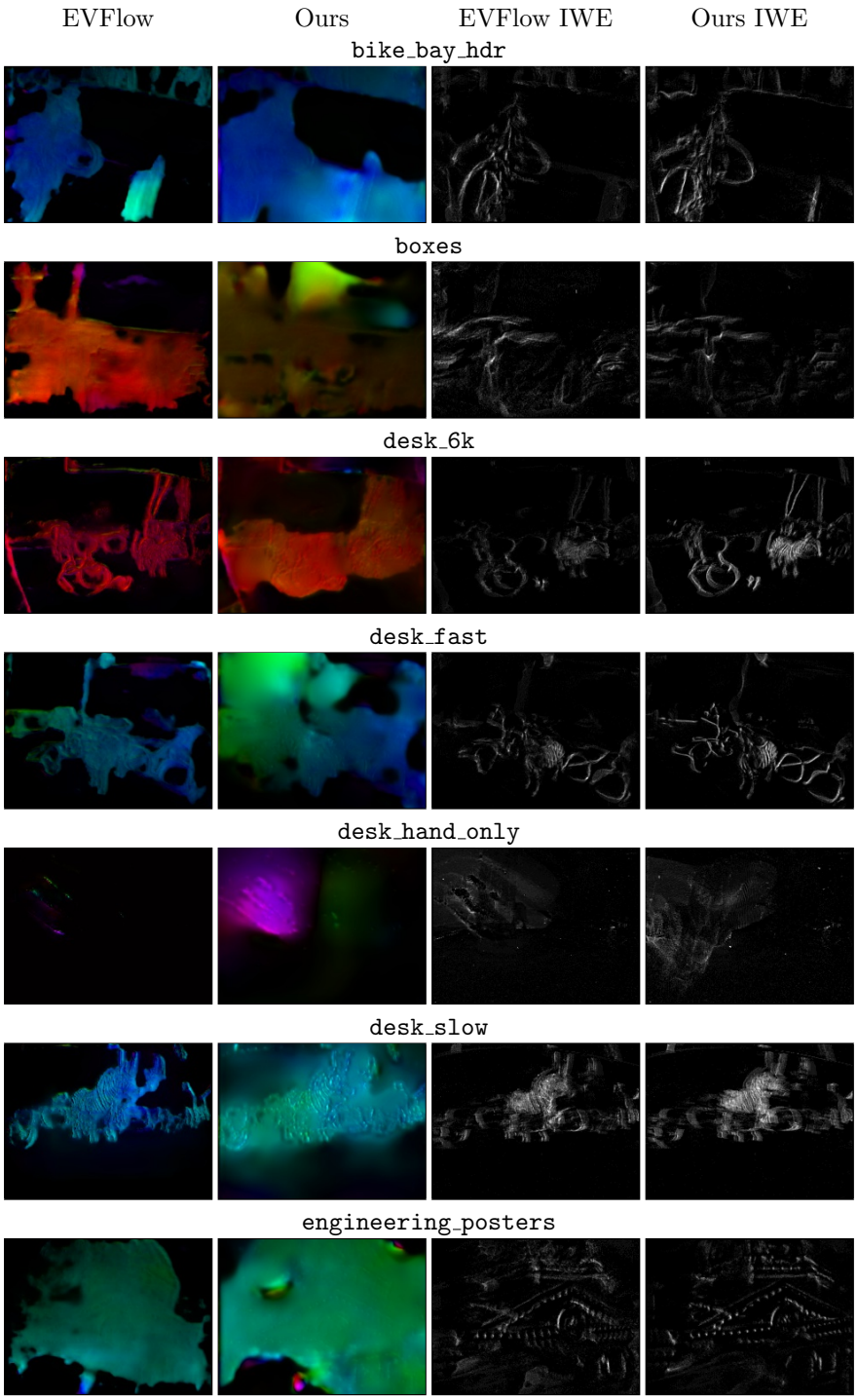


Table 13: Qualitative results for CED [33]. Random selection, not cherry picked. As a matter of interest, the Jenga sequence shows a region of the scene where there is only blank wall, so few events have been generated, resulting in the peculiar artifacts seen in the top left corner.


EVFlow
Ours
EVFlow IWE
Ours IWE
bike_bay_hdr
boxes
desk_6k
desk_fast
desk_hand_only
desk_slow
engineering_posters

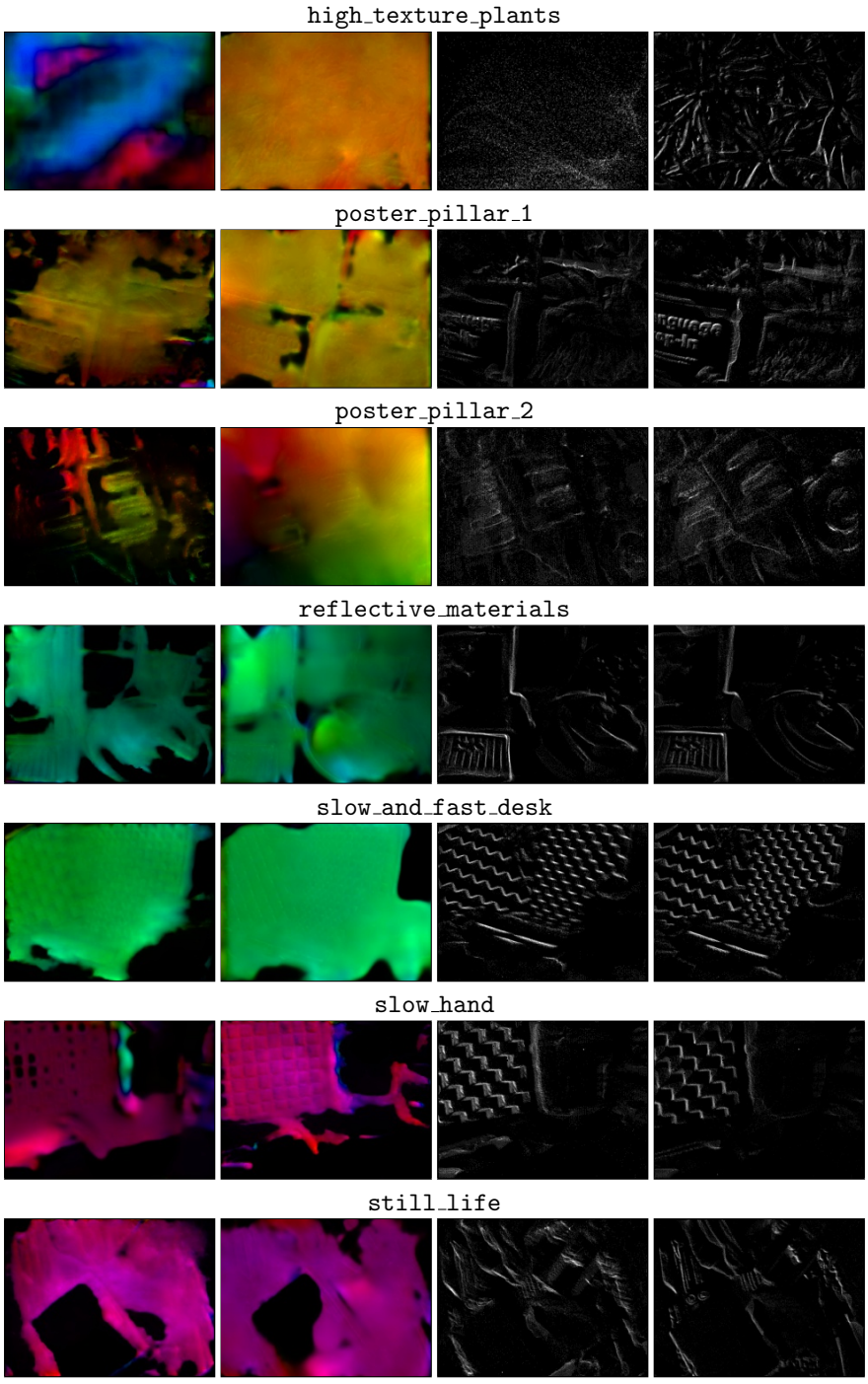
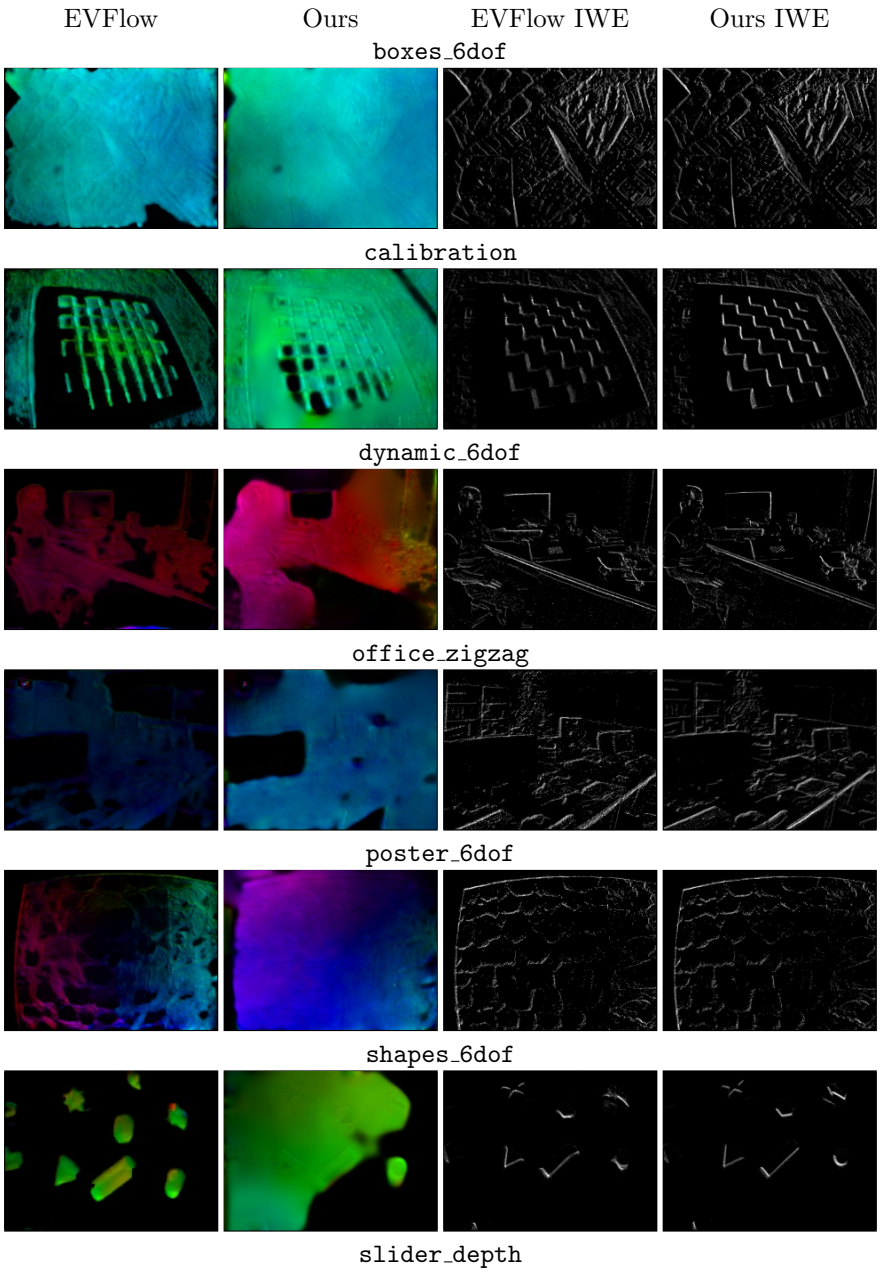

high_texture_plants
poster_pillar_1
poster_pillar_2
reflective_materials
slow_and_fast_desk
slow_hand
still_life

Table 14: Qualitative results for HQFD. Left: optic flow vectors represented in HSV color space, right: image of warped events (IWE). Random selection, not cherry picked.



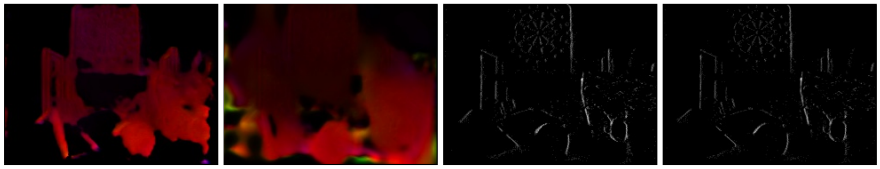
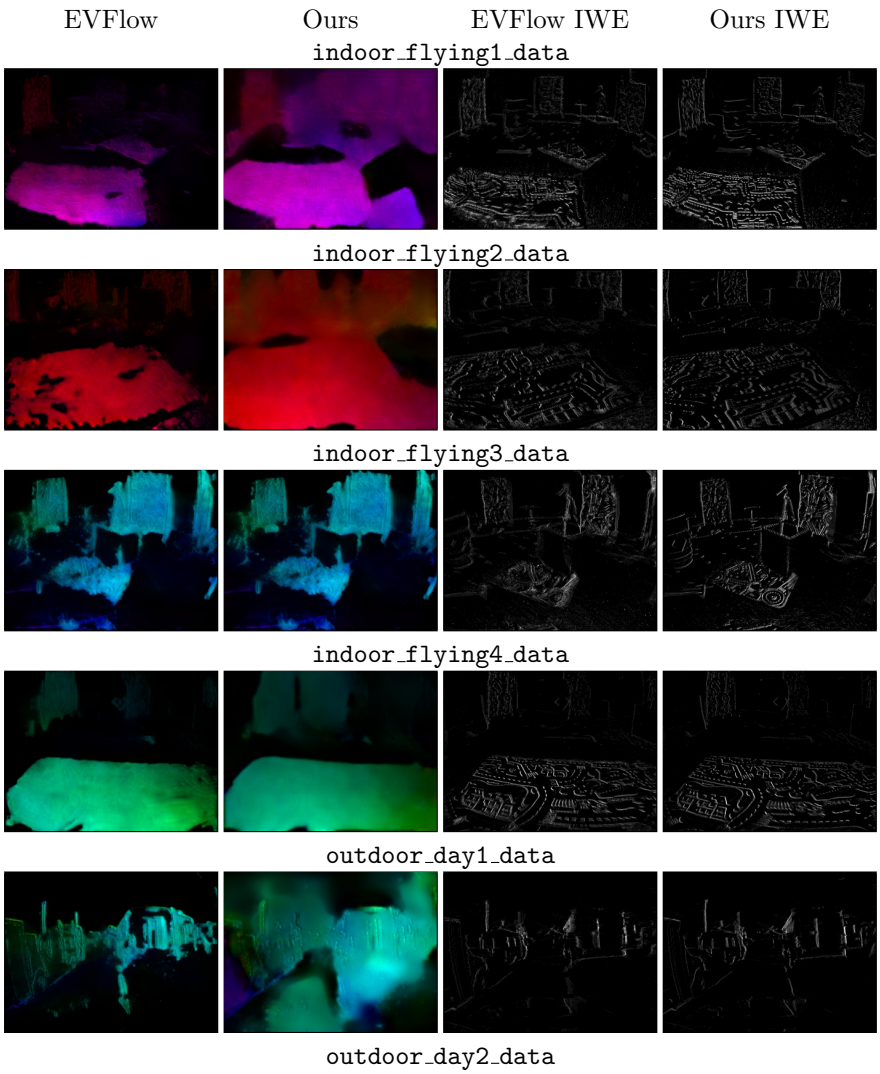


Table 15: Qualitative results for IJRR. Left: optic flow vectors represented in HSV color space, right: image of warped events (IWE). Random selection, not cherry picked.



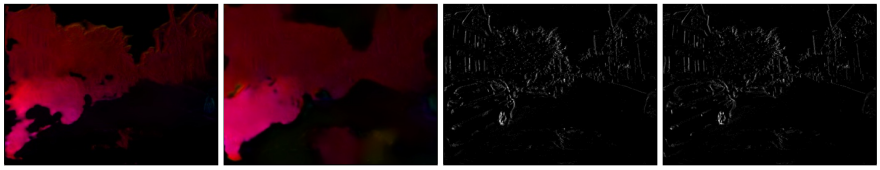


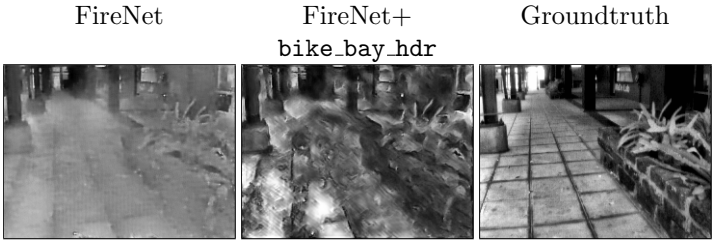
Table 16: Qualitative results for optic flow on MVSEC. Left: optic flow vectors represented in HSV color space, right: image of warped events (IWE). Random selection, not cherry picked.

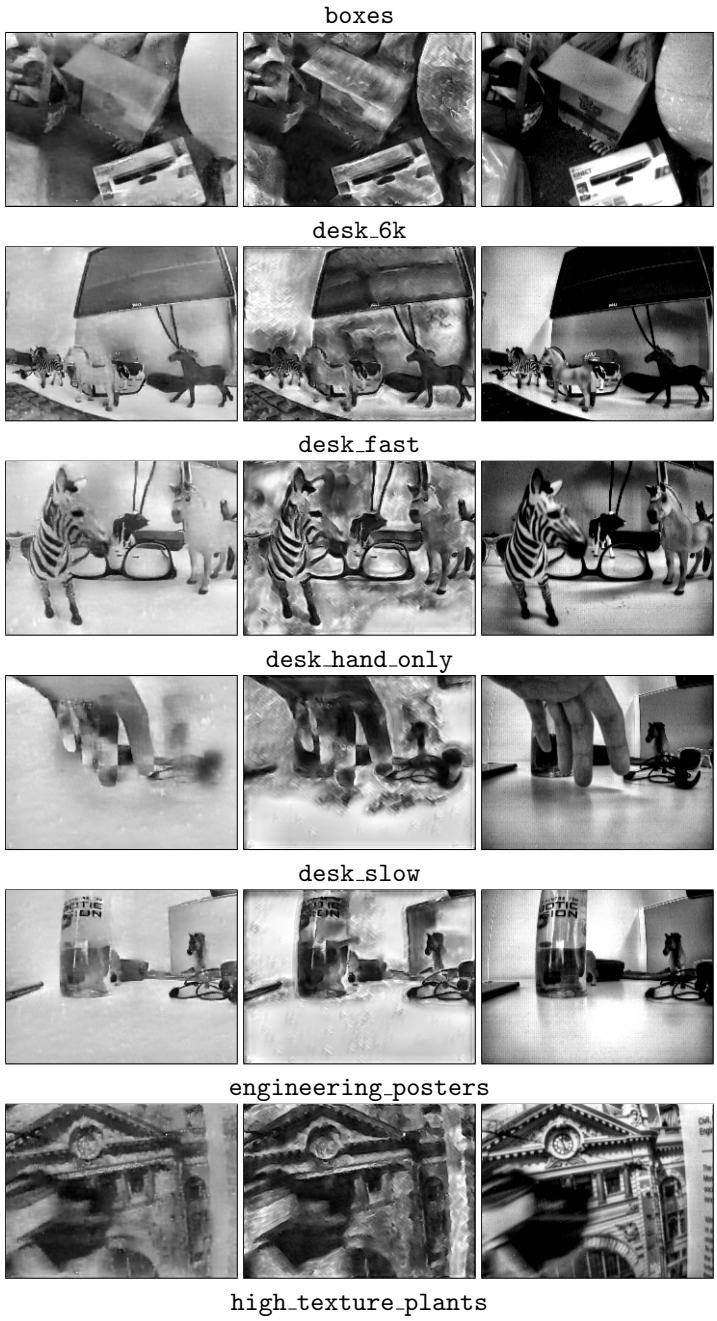
6.7 FireNet

Scheerlinck *et al.* [32] propose a lightweight network architecture for fast image reconstruction with an event camera (FireNet) that has 99.6 % fewer parameters than E2VID [28] while achieving similar accuracy on IJRR [23]. We retrain FireNet using our method and evaluate the original (FireNet) vs. retrained (FireNet+) on IJRR, MVSEC and HQF (Table 17). FireNet+ performs better on HQF and MVSEC though worse on IJRR. One possible explanation is that the limited capacity of a smaller network limits generalizability over a wider distribution of data, and the original FireNet overfits to data similar to IJRR, namely low CTs. If our hypothesis is correct, it presents an additional disadvantage to small networks for event cameras. Comprehensive evaluation (HQF + IJRR + MVSEC) reveals bigger performance gap between FireNet (Table 17) and E2VID (Table 1) architectures than shown in [32] (IJRR only). Qualitatively (Figure 18), FireNet+ looks noisier in textureless regions, while FireNet produces lower contrast images.

Table 17: Mean MSE, SSIM [38] and LPIPS [41] on our HQF dataset, IJRR [23] and MVSEC [42], for original FireNet vs. retrained with our method (FireNet+).

Model	HQF			IJRR			MVSEC		
	MSE	SSIM	LPIPS	MSE	SSIM	LPIPS	MSE	SSIM	LPIPS
FireNet	0.052	**0.514**	0.387	**0.055**	**0.630**	**0.257**	0.182	**0.320**	0.594
FireNet+	**0.049**	0.477	**0.349**	0.058	0.503	0.327	**0.157**	0.288	**0.551**




boxes
desk_6k
desk_fast
desk_hand_only
desk_slow
engineering_posters
high_texture_plants


poster_pillar_1
poster_pillar_2
reflective_materials
slow_and_fast_desk
slow_hand
still_life

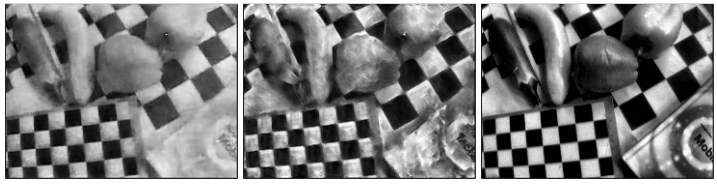
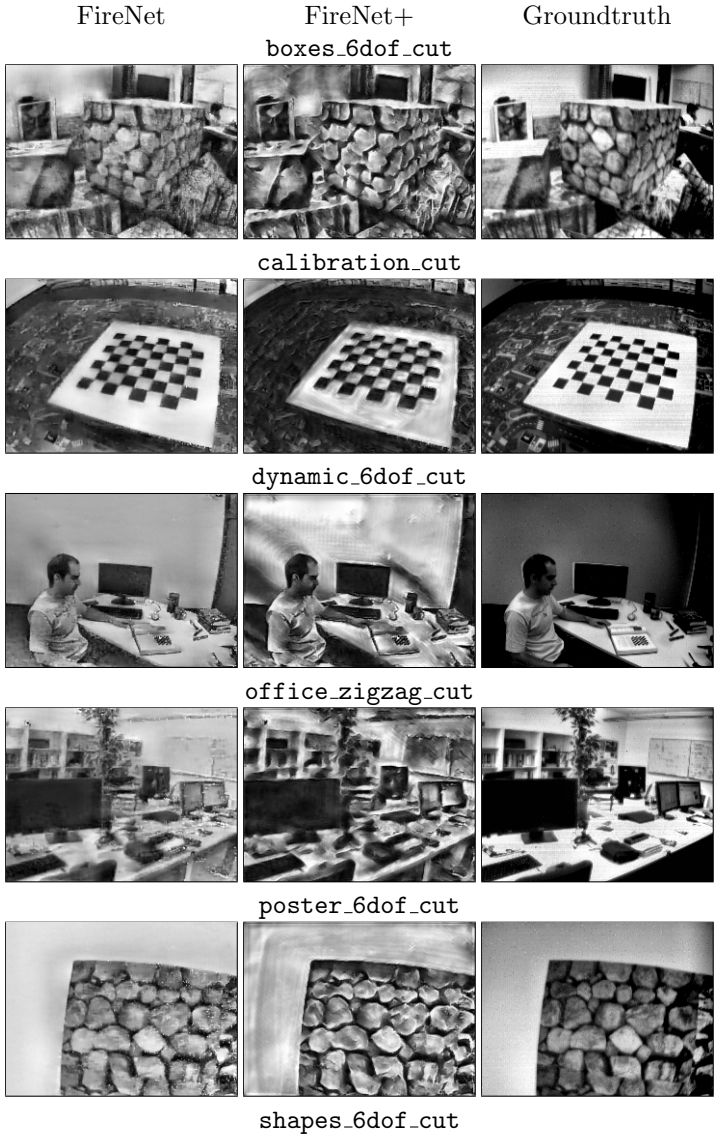


Table 18: Qualitative results for HQFD. Random selection, not cherry picked.



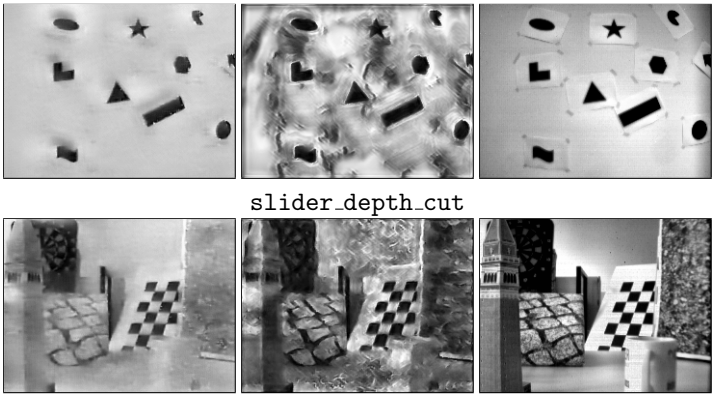
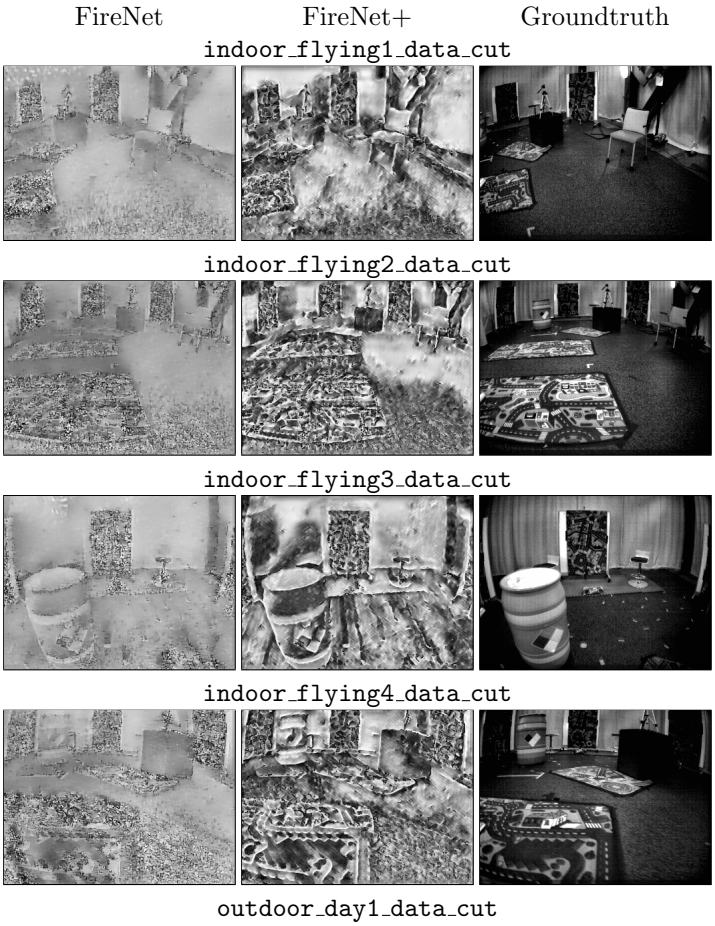


Table 19: Qualitative results for IJRR. Random selection, not cherry picked.



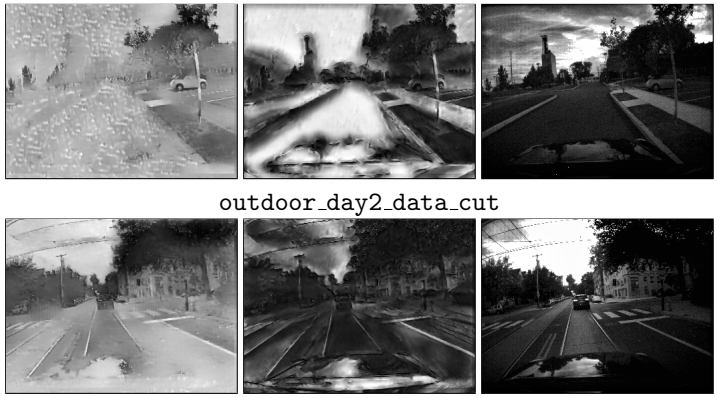


Table 20: Qualitative results for MVSEC. Random selection, not cherry picked.