Supplementary Material for Paper: Event-based Video Reconstruction Using Transformer

Wenming Weng Yueyi Zhang* Zhiwei Xiong
University of Science and Technology of China

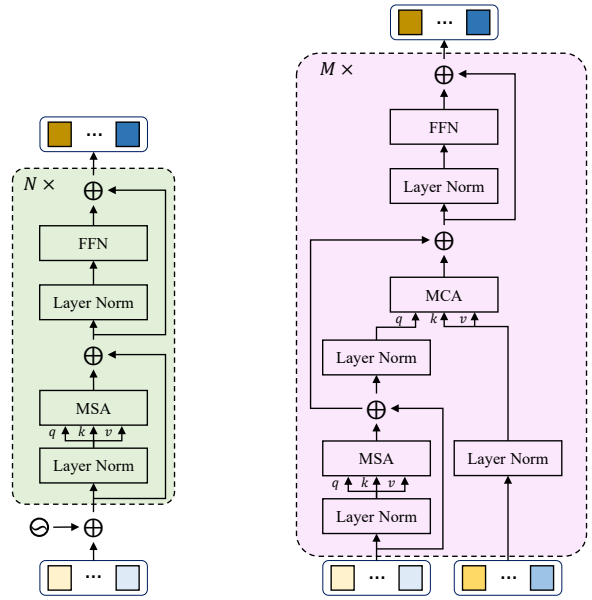
1. Details of network architecture

Table 1 presents the details of our ET-Net architecture. Fig. 1 illustrates the details of Transformer block used in TPA. The embedding dimension of our Transformer Blocks is 256. Eight heads are utilized for MSA/MCA. The dimension setting for the two-layer FFN is 256-1024-256. Additionally, ReLU activation function is adopted after the first linear layer of FFN. In order to alleviate over-fitting, we employ dropout with 0.1 after each MCA/MSA and each linear layer of FFN.

Layer	Description	Output size
Recurrent Convolutional Backbone (RCB)		
Head	Conv2d: $5 \times 5 \times 5 \times 32$, Stride 1, Padding 2 ReLU	$32 \times H \times W$
RB1	Conv2d: $32 \times 5 \times 5 \times 64$, Stride 2, Padding 2 ReLU ConvLSTM	$64 \times \frac{1}{2}H \times \frac{1}{2}W$
RB2	Conv2d: $64 \times 5 \times 5 \times 128$, Stride 2, Padding 2 ReLU ConvLSTM	$128 \times \frac{1}{4}H \times \frac{1}{4}W$
RB3	Conv2d: $128 \times 5 \times 5 \times 256$, Stride 2, Padding 2 ReLU ConvLSTM	$256 \times \frac{1}{8}H \times \frac{1}{8}W$
Token Pyramid Aggregation (TPA)		
TB0-3	Trans-En N + Trans-De M	$\frac{1}{64}HW \times 256$
Multi-Level Upsampler (MLU)		
UB3	Interp2d: upsampling-factor 2 Conv2d: $256 \times 5 \times 5 \times 128$, Stride 1, Padding 2 ReLU	$128 \times \frac{1}{4}H \times \frac{1}{4}W$
UB2	Interp2d: upsampling-factor 2 Conv2d: $128 \times 5 \times 5 \times 64$, Stride 1, Padding 2 ReLU	$64 \times \frac{1}{2}H \times \frac{1}{2}W$
UB1	Interp2d: upsampling-factor 2 Conv2d: $64 \times 5 \times 5 \times 32$, Stride 1, Padding 2 ReLU	$32 \times H \times W$
Tail	Conv2d: $32 \times 1 \times 1 \times 1$, Stride 1, Padding 0 Sigmoid	$1 \times H \times W$

Table 1. Details of the ET-Net architecture. RB, TB and UB denote Recurrent Block, Transformer Block and Upsampling Block, respectively.

*Correspondence should be addressed to zhyuey@ustc.edu.cn



Transformer encoder Transformer decoder

Figure 1. Details of Transformer block. Transformer encoders are responsible for modeling the internal dependency via Multi-head Self-Attention scheme. Transformer decoders are responsible for modeling the intersected dependency via Multi-head Cross-Attention scheme.

2. Sequentialization process in TPA

Fig. 2 shows the illustrative process of sequentialization utilized in TPA. In this diagram as an example, we show features at four scales, whose spatial sizes are 16×16, 8×8, 4×4 and 2×2 respectively. The sizes of patches at four scales are 8×8, 4×4, 2×2 and 1×1 respectively. Therefore, we can produce 4 unfold tokens for each scale in total.

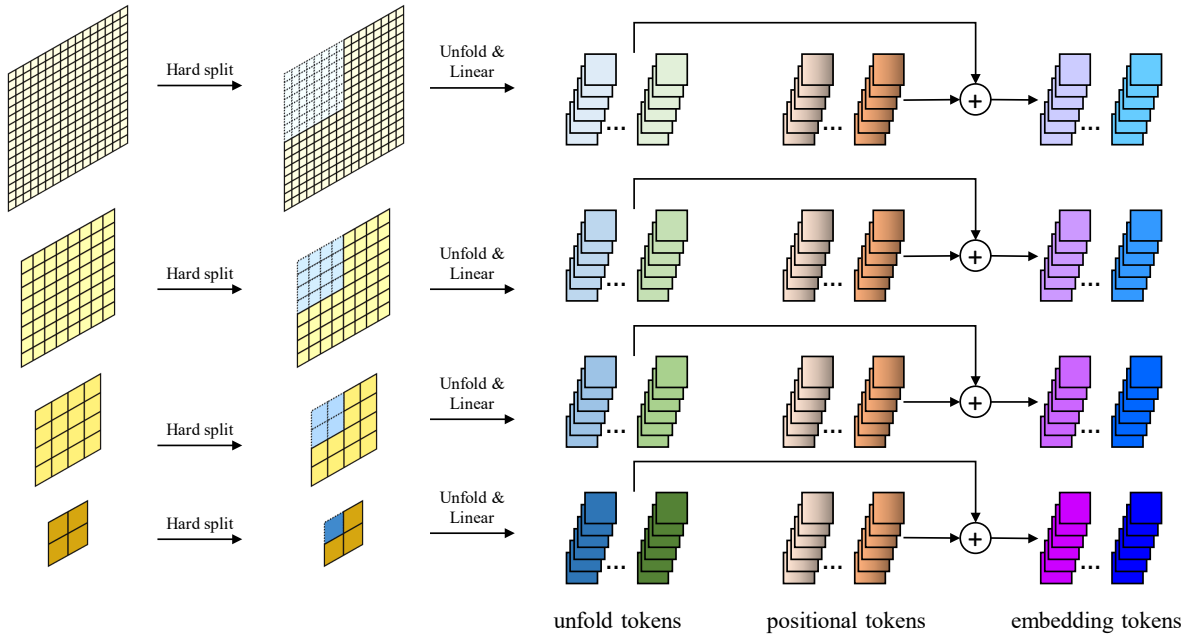


Figure 2. Illustrative process of seuquentialization utilized in TPA.

3. Stacking fashions of Transformer blocks for ET-Net variants

In the main paper, we conduct the ablation study on the aggregation scales in TPA, where we introduce four ET-Net variants. In Fig. 3, we illustrate the stacking fashions of Transformer blocks for the four ET-Net variants. Worthy noting that the ratio of encoder to decoder and stacking structure (square, trapezoid, funnel, etc.) are two factors to form our Transformer blocks. In this work, we apply square staking structure, keep the same ratio and follow the protocol: if the total number is even, we deploy equal encoder/decoder numbers, otherwise the encoder number is one more than the decoder's.

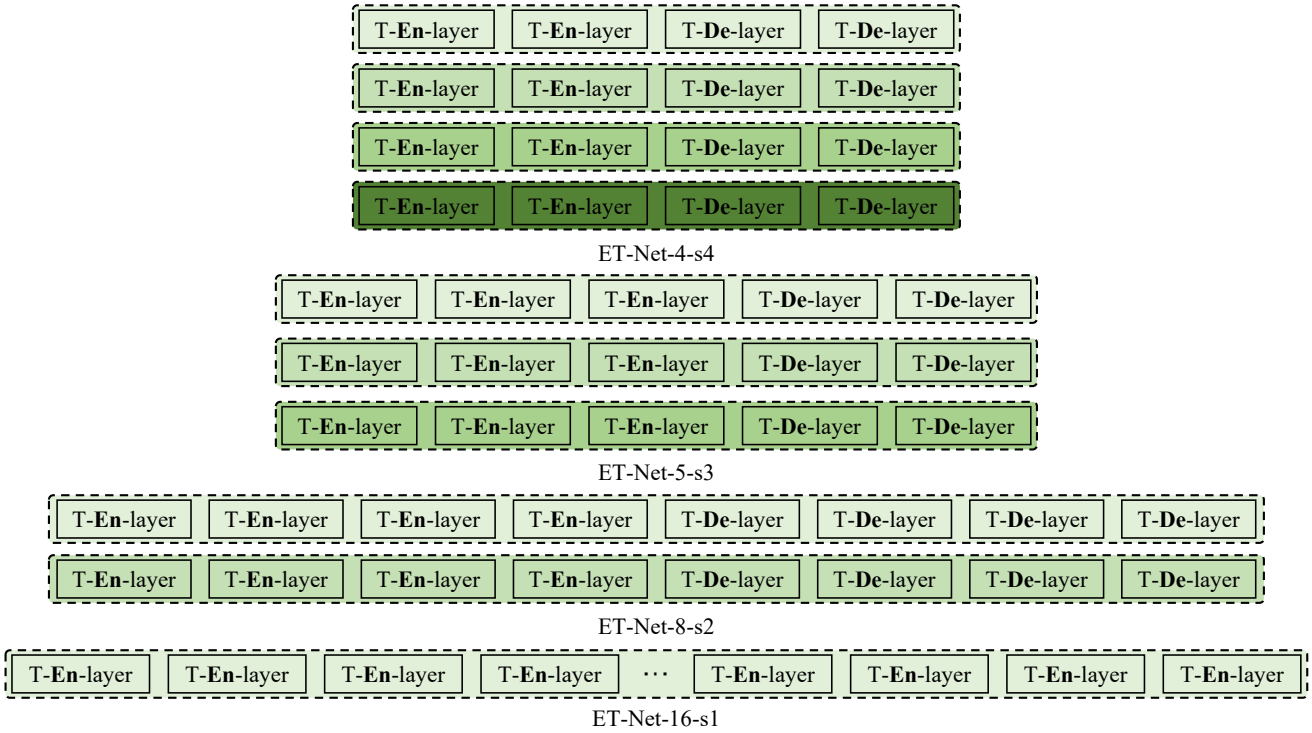


Figure 3. Stacking fashion of Transformer blocks.

4. Sequence cuts

In Table 2, we demonstrate cutting parameters we used for each sequence of the IJRR and MVSEC datasets respectively.

IJRR			MVSEC		
Sequence	Start [s]	End [s]	Sequence	Start [s]	End [s]
boxes_6dof_cut	5.0	20.0	indoor_flying1_data_cut	10.0	70.0
calibration_cut	5.0	20.0	indoor_flying2_data_cut	10.0	70.0
dynamic_6dof_cut	5.0	20.0	indoor_flying3_data_cut	10.0	70.0
office_zigzag_cut	5.0	12.0	indoor_flying4_data_cut	10.0	19.8
poster_6dof_cut	5.0	20.0	outdoor_day1_data_cut	0.0	60.0
shape_6dof_cut	5.0	20.0	outdoor_day2_data_cut	100.0	160.0
slider_depth_cut	1.0	2.5			

Table 2. Sequence cuts for the sequences from IJRR and MVSEC.

5. Breakdown of quantitative results

Table 3 shows the breakdown of the quantitative results of our ET-Net, FireNet+ [3] and E2VID+ [3] on the HQF, IJRR and MVSEC respectively, which are consistent with quantitative results in the main paper. No post-processing procedures are applied for all methods.

Sequences	MSE ↓			SSIM ↑			LPIPS ↓		
	FireNet+	E2VID+	Ours	FireNet+	E2VID+	Ours	FireNet+	E2VID+	Ours
HQF									
bike_bay_hdr	0.0353	0.0362	0.0320	0.586	0.623	0.644	0.354	0.298	0.304
boxes	0.0483	0.0490	0.0403	0.550	0.579	0.603	0.309	0.264	0.246
desk_6k	0.0435	0.0282	0.0385	0.599	0.676	0.662	0.284	0.191	0.218
desk_fast	0.0389	0.0321	0.0345	0.628	0.702	0.690	0.301	0.211	0.237
desk_hand_only	0.0650	0.0447	0.0553	0.657	0.706	0.667	0.425	0.344	0.421
desk_slow	0.0849	0.0375	0.0452	0.645	0.713	0.689	0.311	0.227	0.269
engineering_posters	0.0344	0.0413	0.0390	0.575	0.583	0.600	0.343	0.314	0.299
high_texture_plants	0.0397	0.0264	0.0224	0.561	0.607	0.619	0.182	0.163	0.151
poster_pillar_1	0.0300	0.0316	0.0161	0.579	0.604	0.639	0.345	0.269	0.269
poster_pillar_2	0.0596	0.0264	0.0179	0.583	0.654	0.667	0.391	0.239	0.279
reflective_materials	0.0540	0.0397	0.0548	0.586	0.624	0.613	0.333	0.289	0.303
slow_and_fast_desk	0.0374	0.0407	0.0252	0.614	0.654	0.682	0.309	0.237	0.229
slow_hand	0.0478	0.0477	0.0353	0.542	0.596	0.610	0.397	0.317	0.343
still_life	0.0327	0.0385	0.0321	0.620	0.611	0.622	0.282	0.247	0.270
Mean	0.0465	0.0371	**0.0349**	0.595	0.638	**0.643**	0.326	**0.258**	0.274
IJRR									
boxes_6dof_cut	0.0252	0.0389	0.0140	0.604	0.619	0.692	0.280	0.238	0.243
calibration_cut	0.0248	0.0332	0.0405	0.663	0.639	0.629	0.196	0.187	0.195
dynamic_6dof_cut	0.1410	0.1350	0.1327	0.317	0.298	0.303	0.384	0.352	0.336
office_zigzag_cut	0.0214	0.0420	0.0298	0.507	0.487	0.509	0.298	0.267	0.246
poster_6dof_cut	0.0523	0.0693	0.0521	0.467	0.462	0.531	0.245	0.221	0.221
shape_6dof_cut	0.0858	0.0904	0.0308	0.630	0.755	0.850	0.356	0.186	0.159
slider_depth_cut	0.0467	0.0465	0.0519	0.561	0.599	0.581	0.327	0.239	0.264
Mean	0.0568	0.0650	**0.0503**	0.535	0.551	**0.585**	0.298	0.241	**0.237**
MVSEC									
indoor_flying1_data_cut	0.2246	0.1392	0.1232	0.257	0.345	0.341	0.551	0.522	0.473
indoor_flying2_data_cut	0.2325	0.1540	0.1480	0.243	0.328	0.316	0.554	0.523	0.483
indoor_flying3_data_cut	0.2311	0.1606	0.1250	0.253	0.327	0.344	0.551	0.529	0.461
indoor_flying4_data_cut	0.2613	0.1330	0.1307	0.206	0.354	0.335	0.608	0.512	0.522
outdoor_day1_data_cut	0.2059	0.1272	0.0714	0.285	0.273	0.355	0.622	0.571	0.548
outdoor_day2_data_cut	0.2117	0.0956	0.0811	0.344	0.391	0.457	0.557	0.451	0.458
Mean	0.2278	0.1350	**0.1133**	0.265	0.337	**0.358**	0.574	0.513	**0.491**

Table 3. Breakdown of quantitative results of our proposed ET-Net, FireNet+ and E2VID+ on the HQF, IJRR and MVSEC. Performances on MSE(↓), SSIM(↑) and LPIPS(↓) metrics are reported for each scene. ↑ indicates that the higher value is better while ↓ indicates that the lower value is better. The best is in bold while the second best is with underline.

Variants	HQF			IJRR			MVSEC		
	MSE ↓	SSIM ↑	LPIPS ↓	MSE ↓	SSIM ↑	LPIPS ↓	MSE ↓	SSIM ↑	LPIPS ↓
ET-Net-2-s4	0.0413	0.619	0.288	0.0902	0.515	0.270	0.113	0.376	0.494
ET-Net-4-s4	0.0403	0.635	0.277	0.0522	0.587	0.236	0.118	0.355	0.491
ET-Net-6-s4	0.0430	0.623	0.286	0.0666	0.549	0.250	0.167	0.312	0.538
E2VID-res6	0.0434	0.601	0.314	0.0756	0.545	0.263	0.165	0.319	0.536
E2VID-res12	0.0388	0.610	0.290	0.0687	0.555	0.250	0.169	0.309	0.521
E2VID-res16	0.0416	0.620	0.281	0.0745	0.545	0.256	0.180	0.311	0.518
ET-Net-4-s4	0.0403	0.635	0.277	0.0522	0.587	0.236	0.118	0.355	0.491
ET-Net-5-s3	0.0408	0.629	0.273	0.0584	0.564	0.242	0.136	0.322	0.490
ET-Net-8-s2	0.0387	0.628	0.291	0.0636	0.547	0.260	0.120	0.341	0.510
ET-Net-16-s1	0.0586	0.597	0.310	0.0991	0.509	0.284	0.168	0.305	0.523

Table 4. All ablation results of our proposed ET-Net variants and E2VID variants on the HQF, IJRR and MVSEC. Performances on MSE(↓), SSIM(↑) and LPIPS(↓) metrics are reported for each variants. ↑ indicates that the higher value is better while ↓ indicates that the lower value is better. The best is in bold while the second best is with underline.

6. More ablation results and additional clarifications

Table 4 shows all ablation results of our proposed ET-Net variants and E2VID variants on the HQF, IJRR and MVSEC respectively. Based on these results, we present additional clarifications and discussions.

Domain gap. As shown in Table 4, the domain gap definitely exists among HQF, IJRR and MVSEC. For example, ET-Net-2-s4 and ET-Net-6-s4 perform worse than ET-Net-4-s4 on HQF and IJRR, while ET-Net-4-s4 perform better than the others on MVSEC, which indicates that on HQF and MVSEC, a model with small capacity is not capable of capturing the long range dependency from the latent CNN features, while a large model shows overfitting and degrades the generalization ability. Due to this domain gap, furthermore our network is trained solely on the synthetic training dataset, thus we cannot guarantee the same generalization on all real-world datasets. In the nutshell, we conclude our ablation results in the main paper based on the results on all testing datasets, presenting the overall trend.

Best configuration. In order to search for the best configuration, we conduct experiments with the total number 2 (1 encoder + 1 decoder), 4 (2 encoders + 2 decoders), 6 (3 encoders + 3 decoders). Table 4 shows that the best performance should be achieved near the place where the total number is 4. Then we conduct more experiments with the total number 3 (2 encoder + 1 decoder), 4 (2 encoders + 2 decoders), 5 (3 encoders + 2 decoders). We achieve the best performance when the total number is 5 with three scales as reported in the main paper.

7. High Speed and HDR scenes

We further apply our ET-Net to the High Speed and HDR scenes. Fig. 4 demonstrates that our ET-Net performs well in the High Speed and HDR scenes, recovering more details invisible to conventional cameras.

hdr_tunnel hdr_sun gun_bullet_mug gun_bullet_gnome

Figure 4. Visual results of our ET-Net on the sequences from High Speed and HDR datasets [1].

8. Additional qualitative results

Figs. 5, 6 and 7 show more qualitative comparisons of our ET-Net with baselines (FireNet [2], FireNet+, E2VID [1] and E2VID+) on the sequences from MVSEC, HQF and IJRR, respectively.

9. Reconstructed video clips

The video clips reconstructed from HQF, IJRR and MVSEC datasets using ET-Net as well as other baselines (FireNet, FireNet+, E2VID and E2VID+) are provided in the supplementary file. It should be noted that the video clips don't contain the full span of each sequence. Only a portion of each sequence is used to reconstruct video clips.

References

[1] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[2] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 156–163, 2020.

[3] 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 *European Conference on Computer Vision*. Springer, 2020.

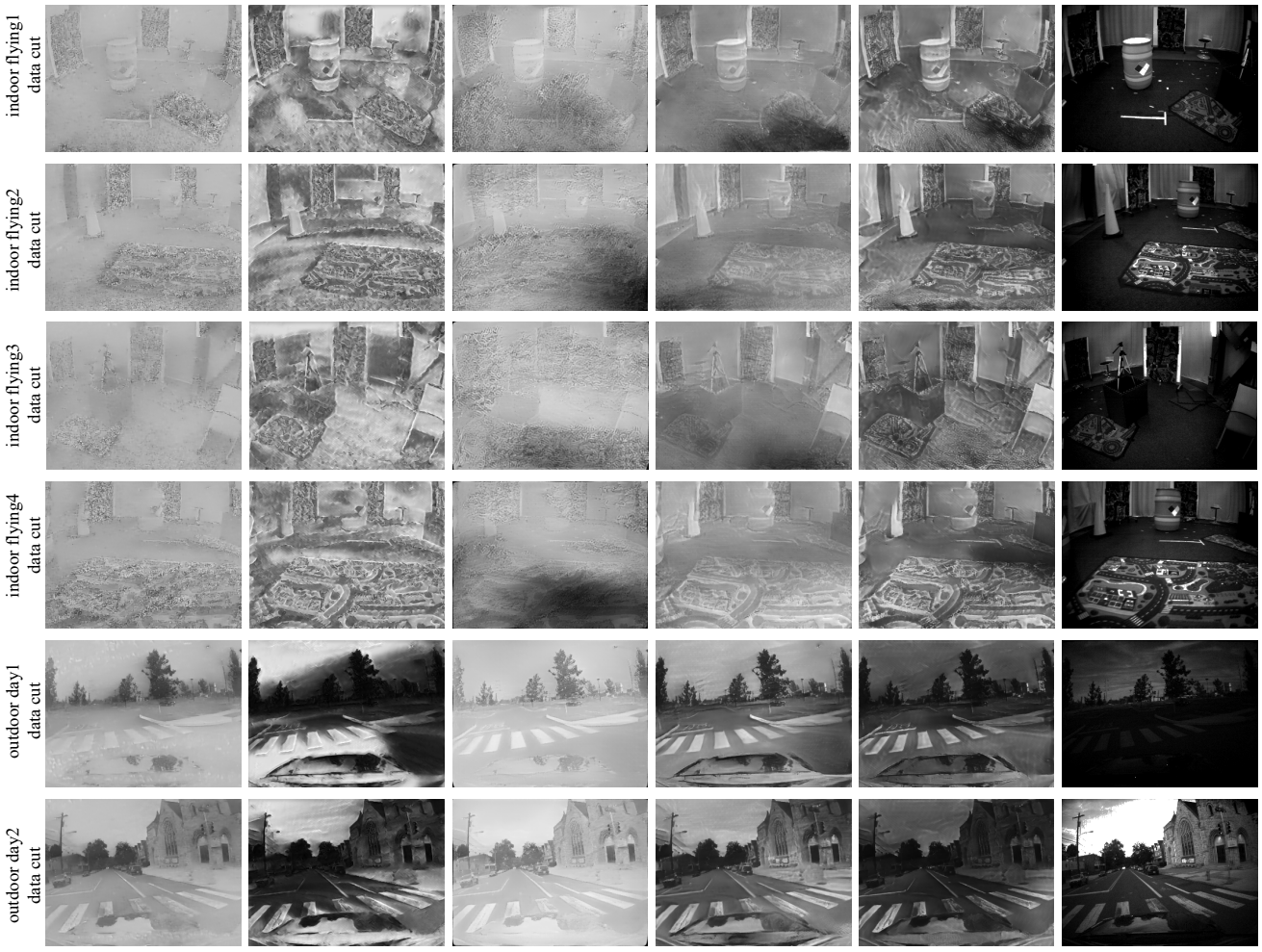
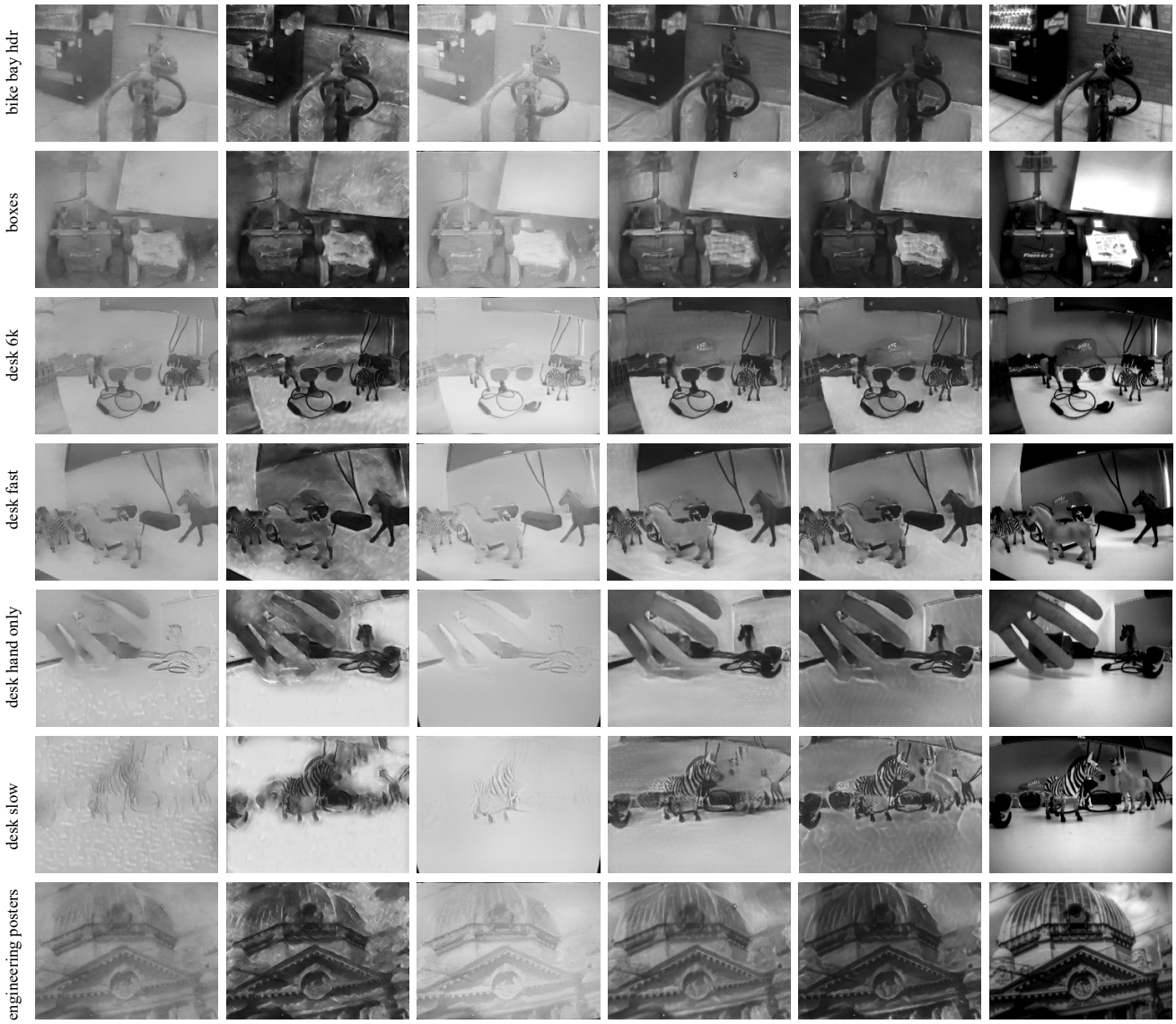


Figure 5. Additional qualitative results of our ET-Net, FireNet, FireNet+, E2VID and E2VID+ on the sequences from MVSEC dataset.


bike bay hdr
boxes
desk 6k
desk fast
desk hand only
desk slow
engineering posters
FireNet
FireNet+
E2VID
E2VID+
ET-Net (Ours)
Ground-truth

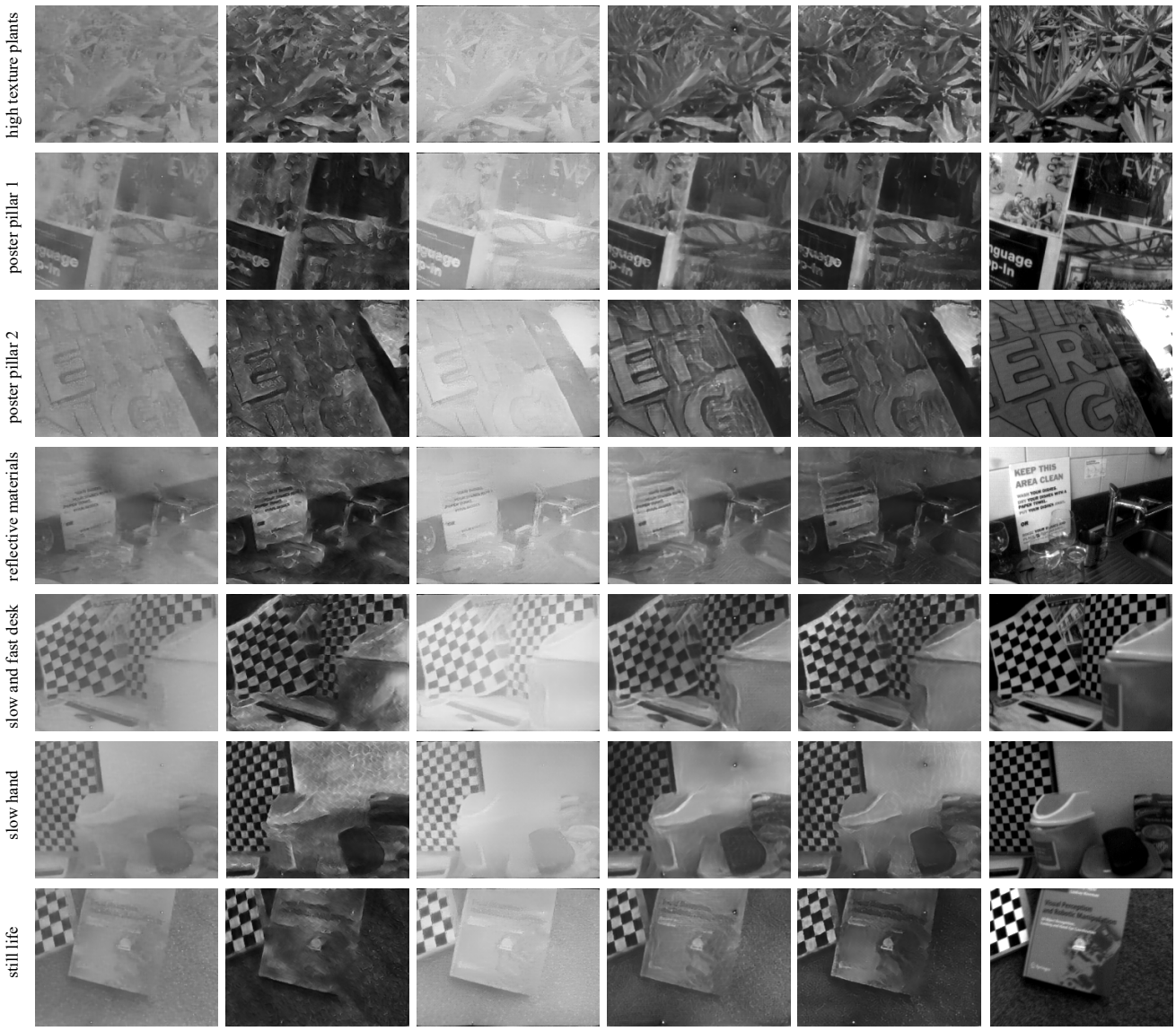


Figure 6. Additional qualitative results of our ET-Net, FireNet, FireNet+, E2VID and E2VID+ on the sequences from HQF dataset.

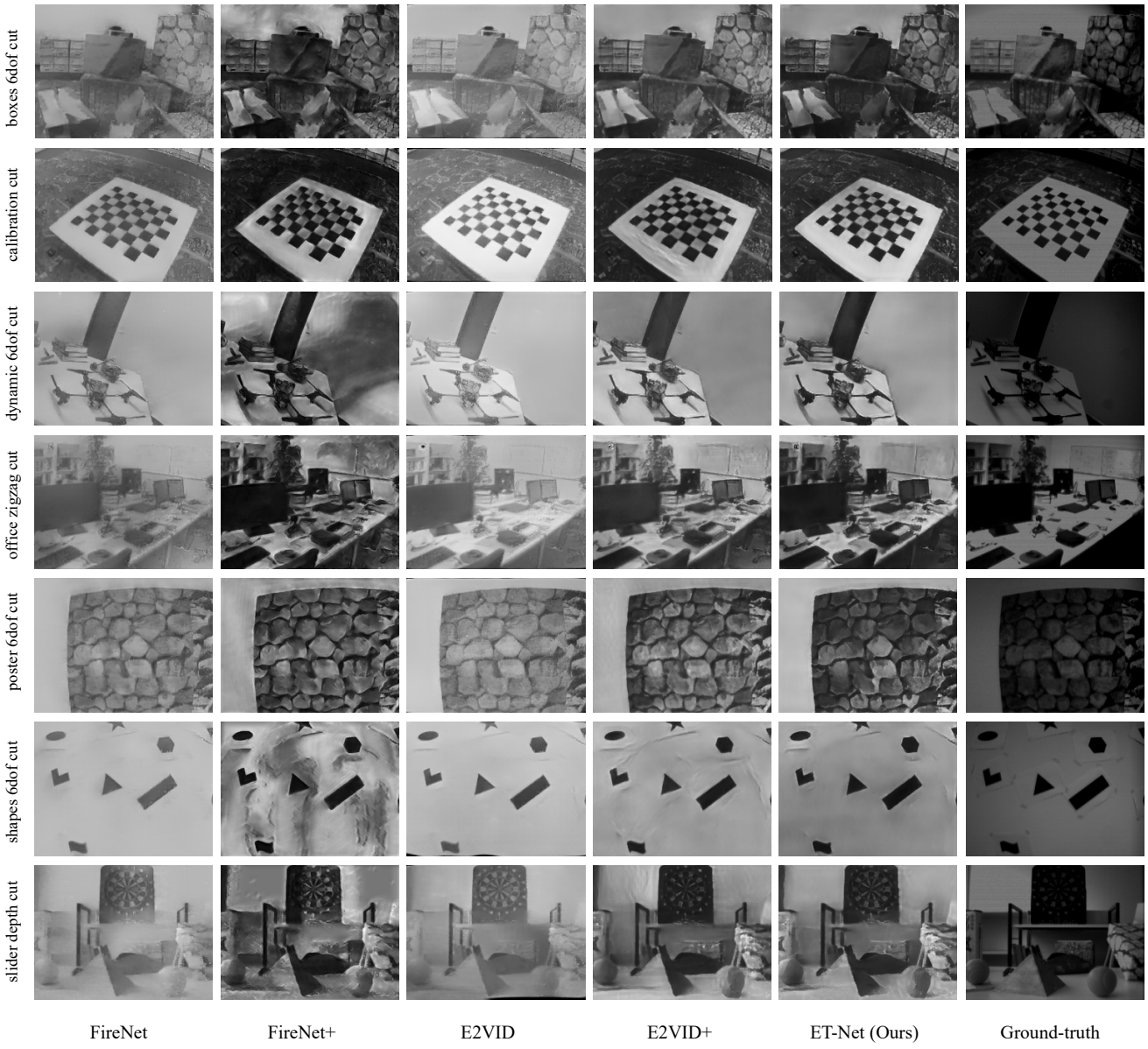


Figure 7. Additional qualitative results of our ET-Net, FireNet, FireNet+, E2VID and E2VID+ on the sequences from IJRR dataset.