Supplementary Material - Multi-Modal Fusion of Event and RGB for Monocular Depth Estimation Using a Unified Transformer-based Architecture

This supplementary material supports the findings presented in our paper. Due to space limitations in the main text, additional results on the outdoor night2 and outdoor night3 datasets are provided here. These tables offer a more comprehensive evaluation of our model's performance across different environmental conditions.

Section 4 shows the results of our unified transformer model on MVSEC and EventScape Datasets. MVSEC consists of four driving datasets, Outdoor night1, night2, night3 and day1. In the main text, we tabulated the results on night1 and day1 to show the model's performance in different lightning conditions. Table 2 tabulates the different metric results of our model with HMNet [1] on outdoor night2 and night3 datasets.

Table 3 shows the ablation study on night2 and night3 datasets for components impact analysis and transformer encoder analysis.

References

[1] Ryuhei Hamaguchi, Yasutaka Furukawaa, Masaki Onishi, and Ken Sakurada. Hierarchical neural memory network for low latency event processing. *CVPR*, 2023. 1, 2

Experiments	Outdoor Night 2			Outdoor Night3		
	10m (↓)	20m (↓)	30m(↓)	10m (↓)	20m (↓)	30m (↓)
Transformer-based (best)	**1.54**	**2.23**	**2.95**	**1.24**	**1.96**	**2.81**
Without the convLSTM	2.51	3.01	3.6	2.43	2.88	3.45
Without the skip connections	2.17	2.78	3.61	2.08	2.71	3.61

Table 1. Absolute Mean Depth Error on Outdoor Night2 and Night3 for components impact analysis

Metrics	Outdoor Night2						Outdoor Night3					
	HMNet [1]			Our transformer			HMNet [1]			Our transformer		
	10m	20m	30m	10m	20m	30m	10m	20m	30m	10m	20m	30m
Abs Rel (↓)	**0.231**	**0.246**	**0.256**	0.256	0.254	0.260	0.208	0.228	0.240	**0.194**	**0.201**	**0.215**
RMSE (↓)	**2.381**	3.503	**4.482**	2.660	**3.456**	4.491	2.091	3.325	4.391	**2.005**	**2.923**	**2.808**
$\delta < 1.25^1$ (↑)	0.737	0.646	0.607	**0.754**	**0.682**	**0.628**	0.734	0.639	0.597	**0.763**	**0.698**	**0.634**
$\delta < 1.25^2$ (↑)	0.884	0.844	0.817	**0.886**	**0.883**	**0.852**	0.888	0.846	0.818	**0.914**	**0.911**	**0.872**
$\delta < 1.25^3$ (↑)	**0.949**	0.934	0.920	0.941	**0.953**	**0.947**	0.957	0.941	0.927	**0.967**	**0.975**	**0.9653**

Table 2. Different Metric Results on MVSEC Dataset for Outdoor Night2 and Outdoor Night3 Sequences

Experiments	Outdoor Night 2			Outdoor Night3			Model Size(MB)(↓)	Model Parameters(M)(↓)
	10m (↓)	20m (↓)	30m(↓)	10m (↓)	20m (↓)	30m (↓)		
Transformer-based (ours)	**1.54**	**2.23**	**2.95**	**1.24**	**1.96**	**2.81**	**336.37 MB**	**88 M**
Individual Encoders	2.83	4.59	5.06	2.58	4.49	5.01	660.31 MB	173 M
Cross-attention encoders	8.1	6.14	6.39	7.8	5.79	6.13	606.24 MB	158 M

Table 3. Absolute Mean Depth Error on Outdoor Night2 and Night3 dataset along with the Model Size and Number of Model Parameters for transformer encoder analysis