北京大学计算机学院
School of Computer Science

Camera Intelligence
A Computational Photography Lab @ PKU
http://camera.pku.edu.cn

NEURAL INFORMATION PROCESSING SYSTEMS

Dense Metric Depth Estimation via Event-based Differential Focus Volume Prompting

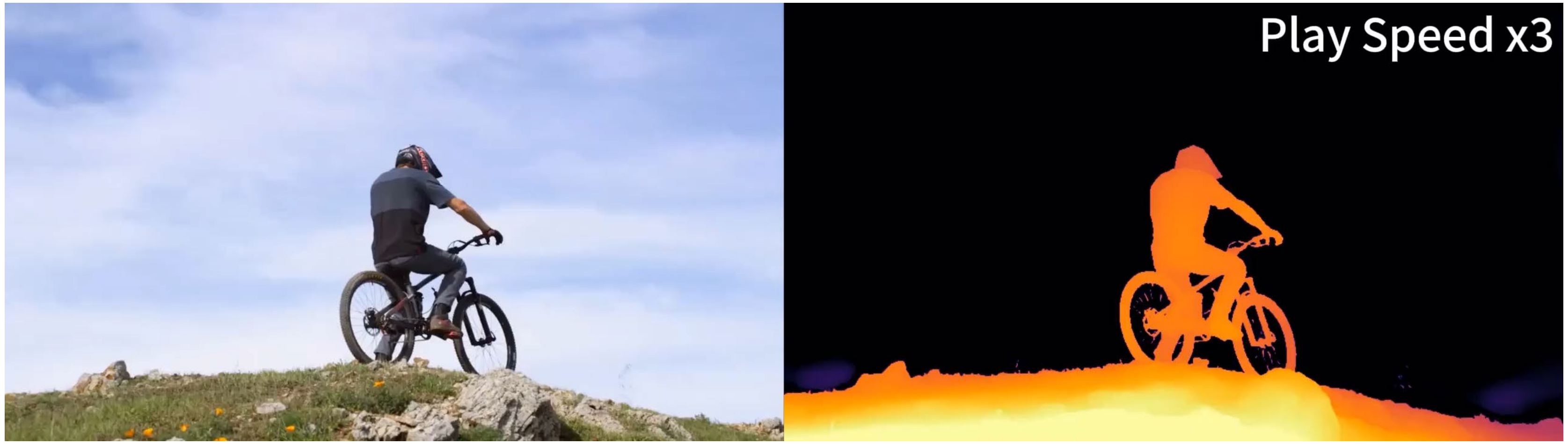
Boyu Li Peiqi Duan* Zhaojun Huang Xinyu Zhou Yifei Xia Boxin Shi*

Peking University

{liboyu, duanqi0001, huangzhaojun, zhouxiny, yfxia, shiboxin}@pku.edu.cn

Depth Estimation

- Applications: 3D Modelling, Autonomous driving, Robotics
- Metric Depth: Absolute depth values of valid pixels
- Relative Depth: Relative depth values normalized to Min & Max



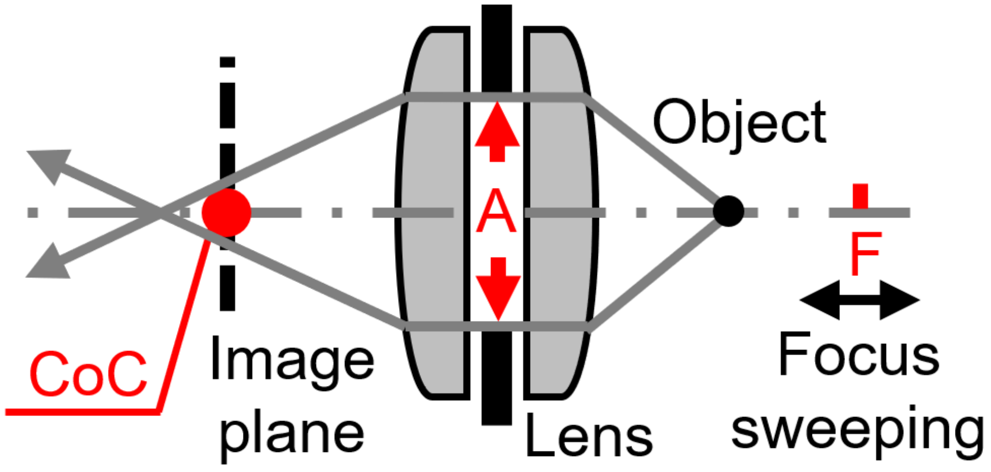
😀 Object shapes

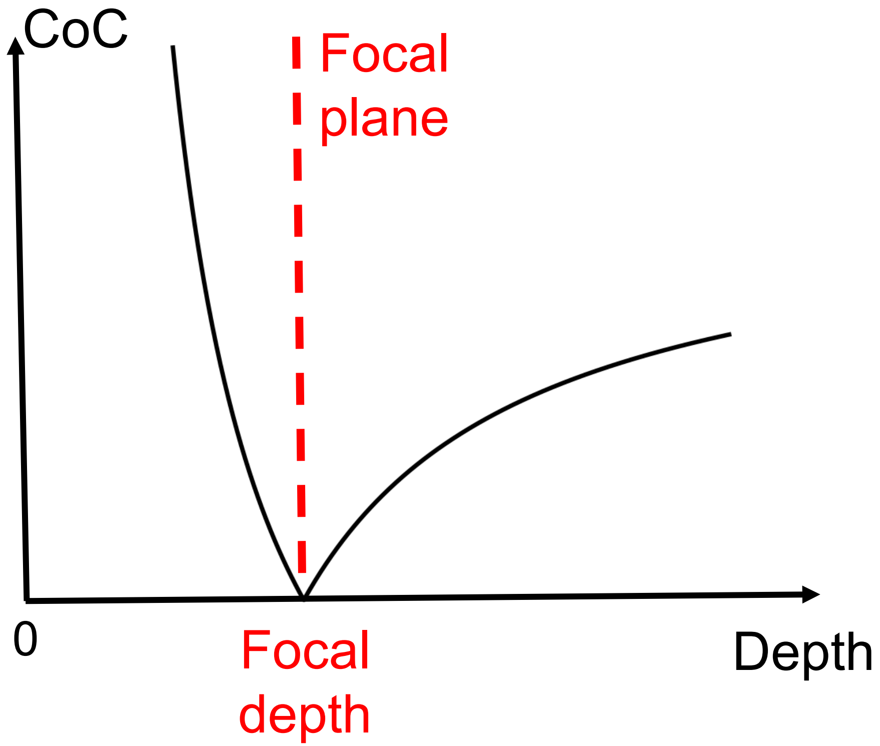
🙁 Absolute values

[1] Chen et al. Video Depth Anything: Consistent Depth Estimation for Super-Long Videos. CVPR 2025.

Depth from Focus

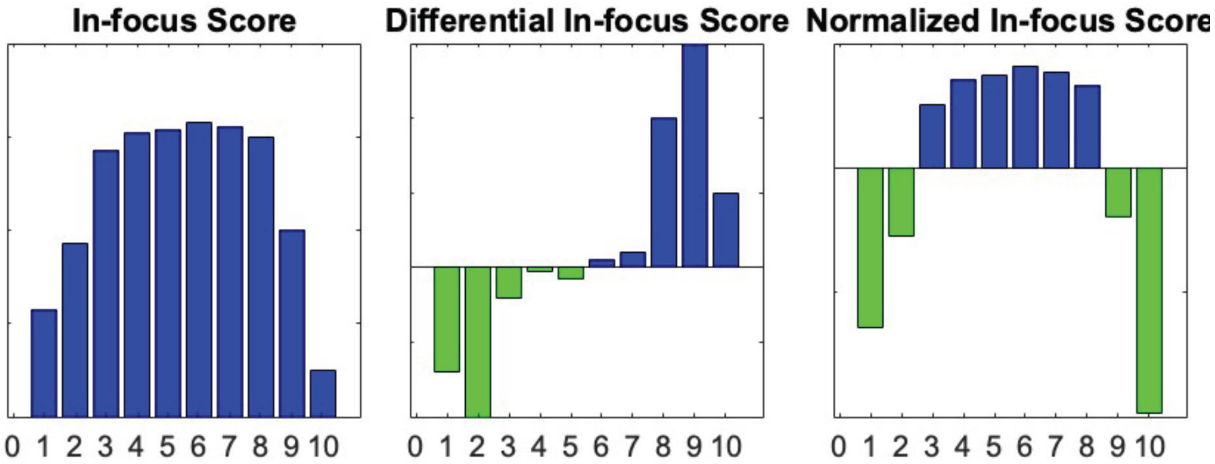
- During focus sweeping, there is an optimal focusing timestamp for each point of the scene, where the Circle of Confusion (CoC) is the smallest.





Depth from Focus

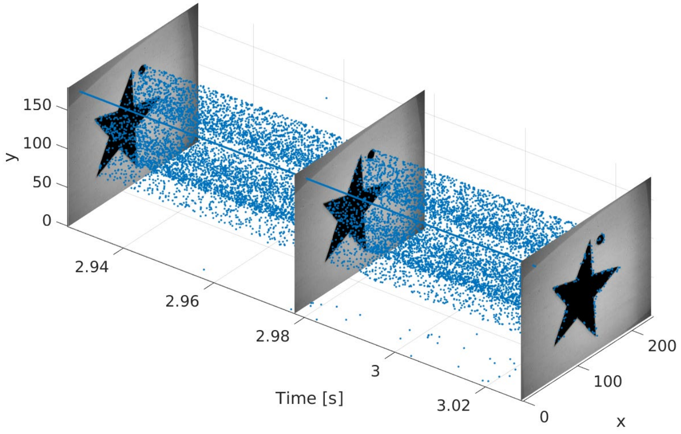
- Estimate depth of a scene by using the information acquired through the change of the focus of a camera
- Focus Volume: Store the in-focus degree of each pixel
- Differential Focus Volume: First-order derivative of Focus Volume



[2] Yang et al. Deep Depth from Focus with Differential Focus Volume. CVPR 2022.

Event-based Vision

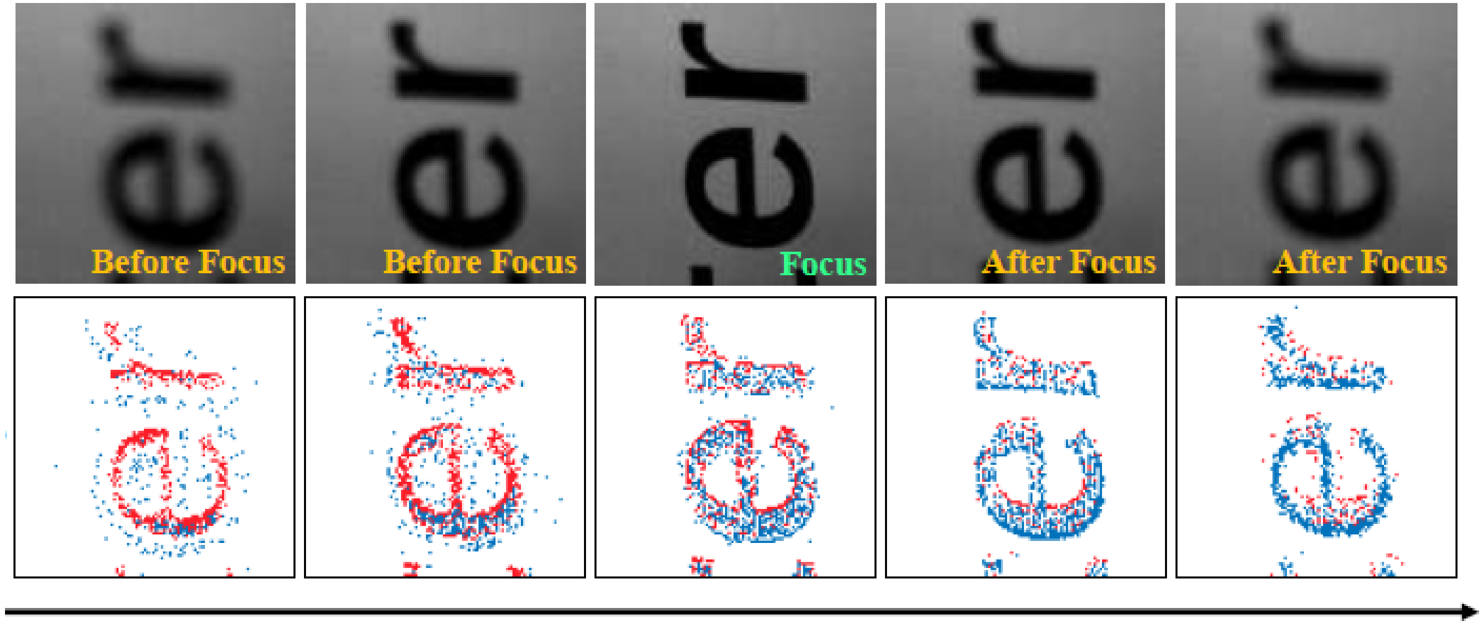
- Traditional cameras can only capture discrete frames with fixed frame rate
- Event cameras can record continuous changing of the scene with asynchronous timestamps



[3] Mueggler et al. The Event-Camera Dataset and Simulator: Event-based Data for Pose Estimation, Visual Odometry, and SLAM. IJRR 2017.

Motivation

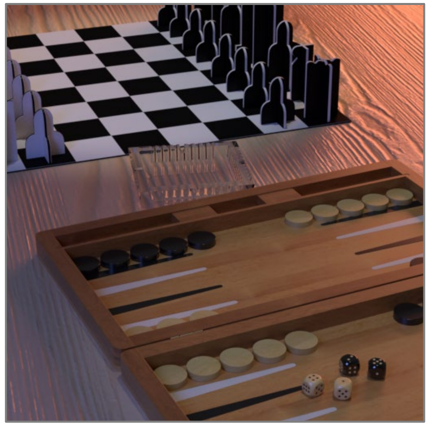
- Events triggered around the intensity-changing pixels of an image may experience a polarity reversal before and after focusing.
- Event-based Differential Focus Volume (EDFV)



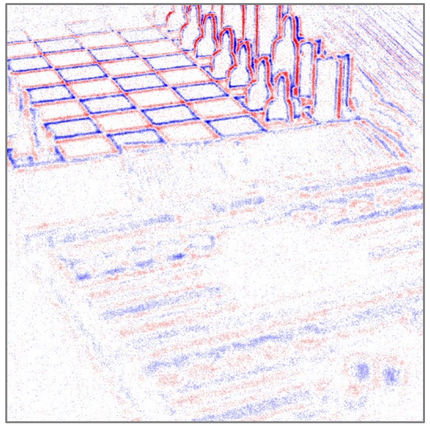
[4] Jiang et al. Learning depth from focus with event focal stack. IEEE SENSORS JOURNAL 2024.

Motivation

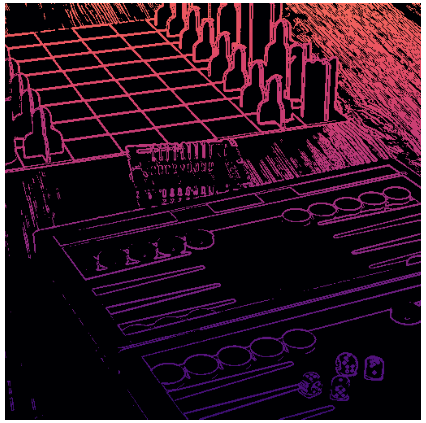
- The sparsity of events makes it hard to get dense predictions.
- In contrast, DFF methods could extract dense information from images.



Image



Events

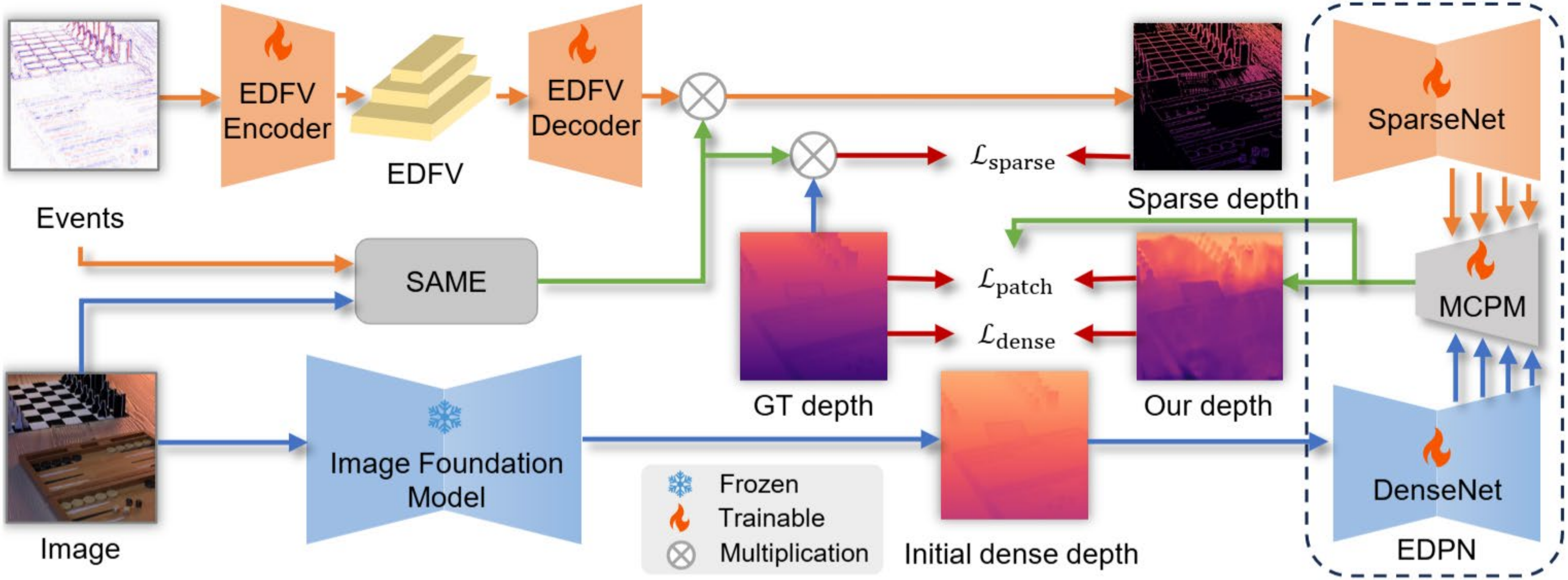


Sparse depth

😀 Absolute values

🙁 Dense prediction

Method



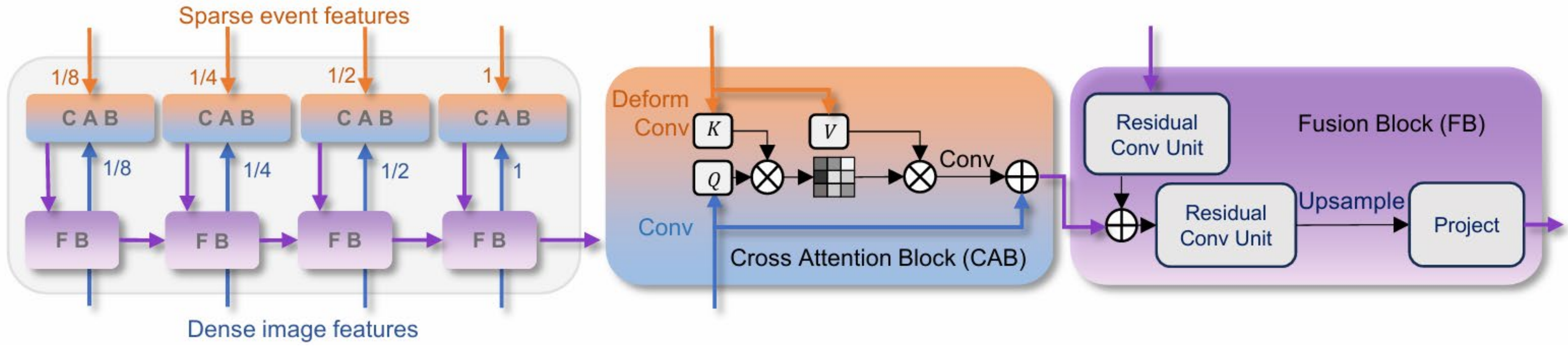
Event-based Depth Prompting Network (EDPN)

Method

- Spatial Attention Mask Extraction (SAME)

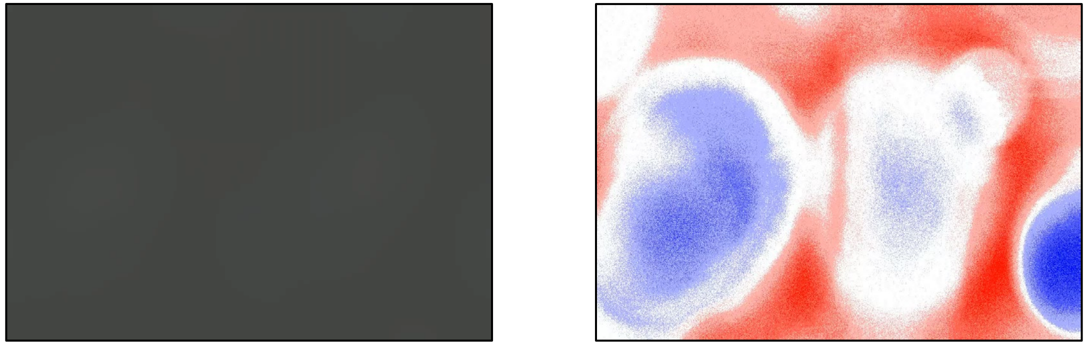
$$\mathbf{M} = \mathrm{Dilate}((\underset{\substack{\uparrow \\ \text{Image} \\ \text{gradients}}}{\nabla \mathbf{I}} > \epsilon_I) \cdot (\underset{\substack{\uparrow \\ \text{Event} \\ \text{density}}}{\rho_e} > \epsilon_e))$$

- Multi-scale Cross-attention-guided Prompting Module (MCPM)

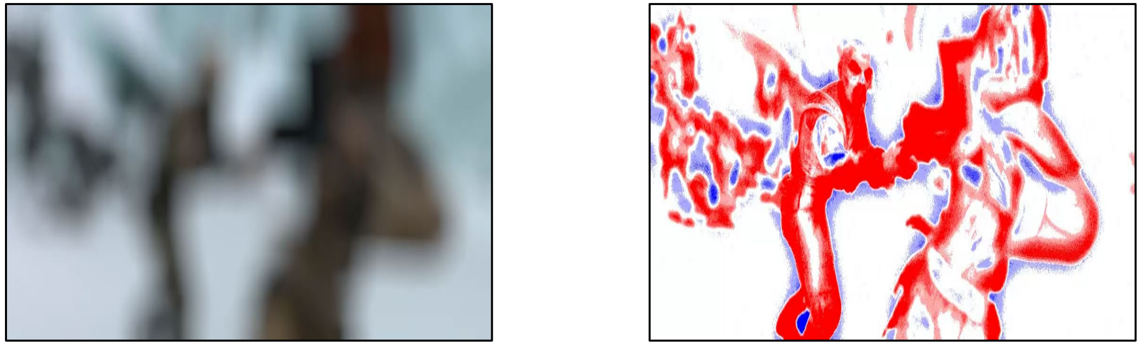


Datasets

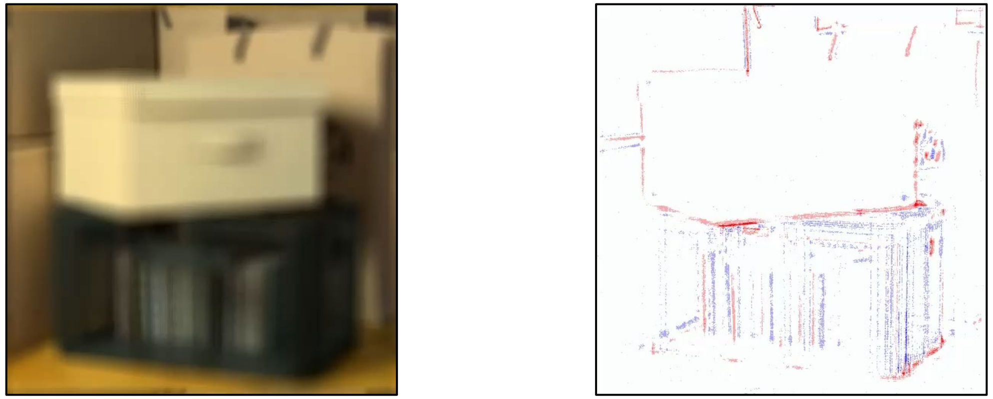
- Blender-Syn



- Sintel-Dr. Bokeh



- 4DLFD-Semi-Real



- EDFV-Real

Results: In-domain Experiments

Method	Type	Blender-Syn						Sintel-Dr. Bokeh					
		RMSE(↓)	AbsRel(↓)	log10(↓)	δ_1(↑)	δ_2(↑)	δ_3(↑)	RMSE(↓)	AbsRel(↓)	log10(↓)	δ_1(↑)	δ_2(↑)	δ_3(↑)
DefocusNet	DFF	0.243	0.372	0.107	0.734	0.818	0.861	0.209	0.728	0.192	0.412	0.644	0.797
DFF-FV	DFF	0.184	0.223	0.062	0.862	0.907	0.926	0.160	0.661	0.109	0.766	0.863	0.898
DFF-DFV	DFF	0.186	0.250	0.062	0.871	0.906	0.923	0.134	0.569	0.104	0.738	0.861	0.907
DDFS	DFF	0.244	0.387	0.109	0.723	0.804	0.849	0.282	1.072	0.282	0.441	0.578	0.648
HybridDepth	DFF	0.089	0.123	0.051	0.823	0.925	0.969	0.273	0.657	0.295	0.233	0.393	0.540
DA V2	Mono	**0.063**	0.089	0.035	0.865	0.956	**0.989**	0.297	0.482	0.361	0.330	0.419	0.472
Metric3D V2	Mono	0.095	0.162	0.062	0.826	0.934	0.973	0.170	0.479	0.174	0.452	0.561	0.754
DAC	Mono	0.176	0.238	0.115	0.654	0.868	0.947	0.273	0.951	0.289	0.268	0.409	0.573
Ours	DFF	0.068	**0.077**	**0.028**	**0.919**	**0.972**	0.987	**0.095**	**0.141**	**0.072**	**0.806**	**0.901**	**0.945**

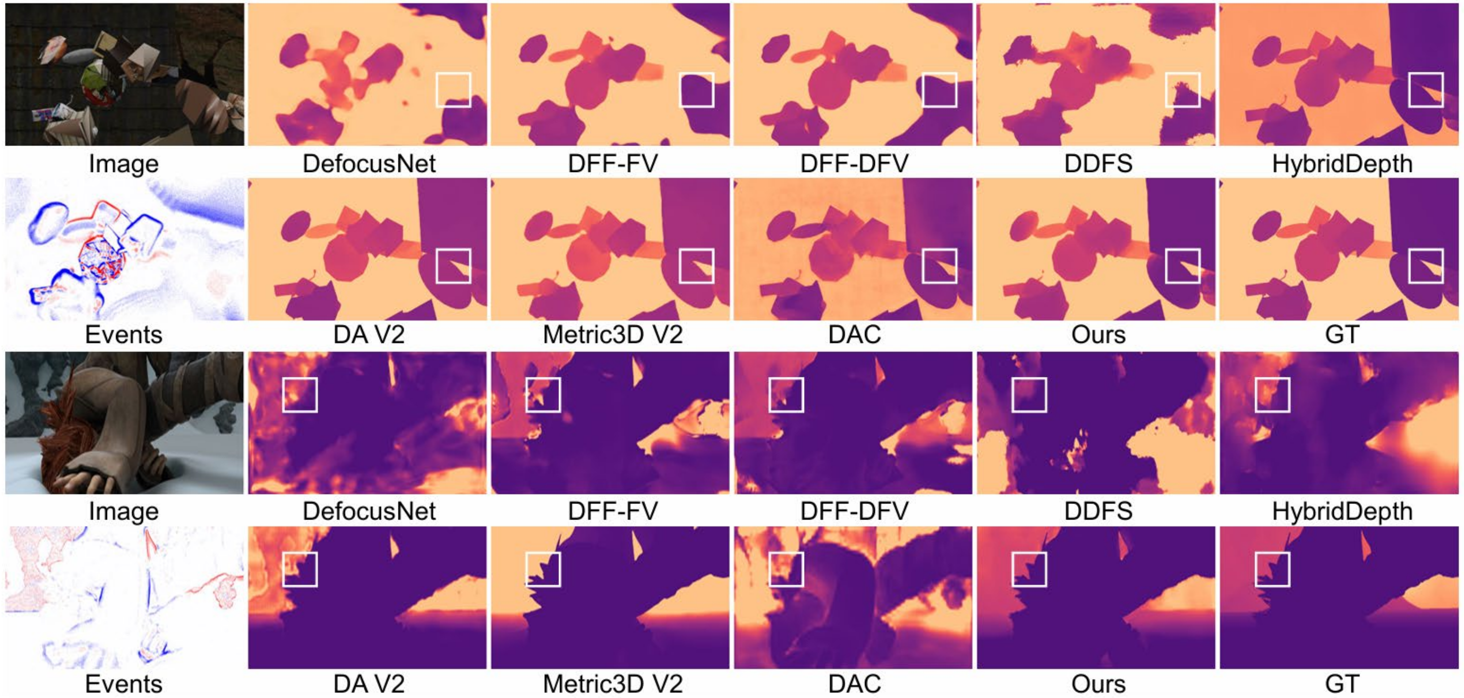


Image DefocusNet DFF-FV DFF-DFV DDFS HybridDepth

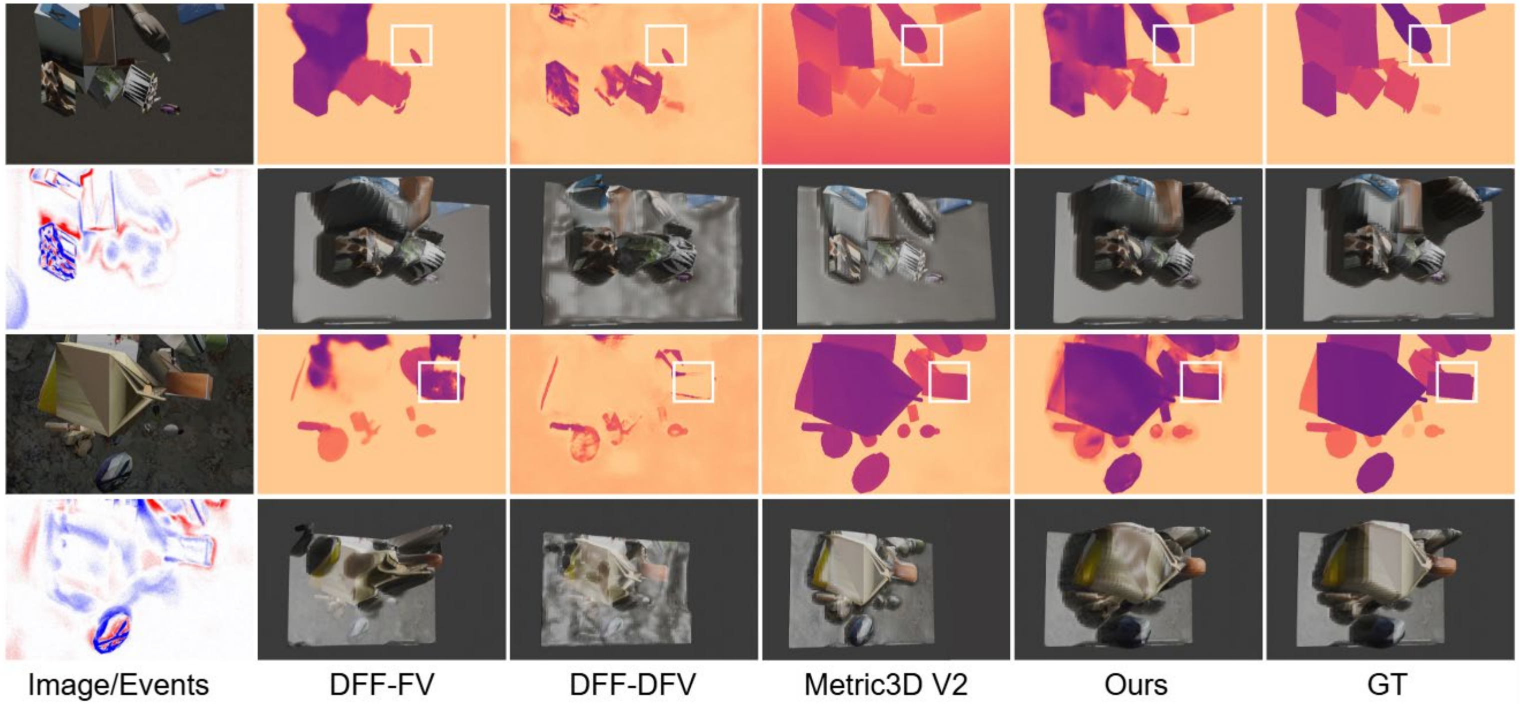
Events DA V2 Metric3D V2 DAC Ours GT

Image DefocusNet DFF-FV DFF-DFV DDFS HybridDepth

Events DA V2 Metric3D V2 DAC Ours GT

Results: Zero-shot Experiments

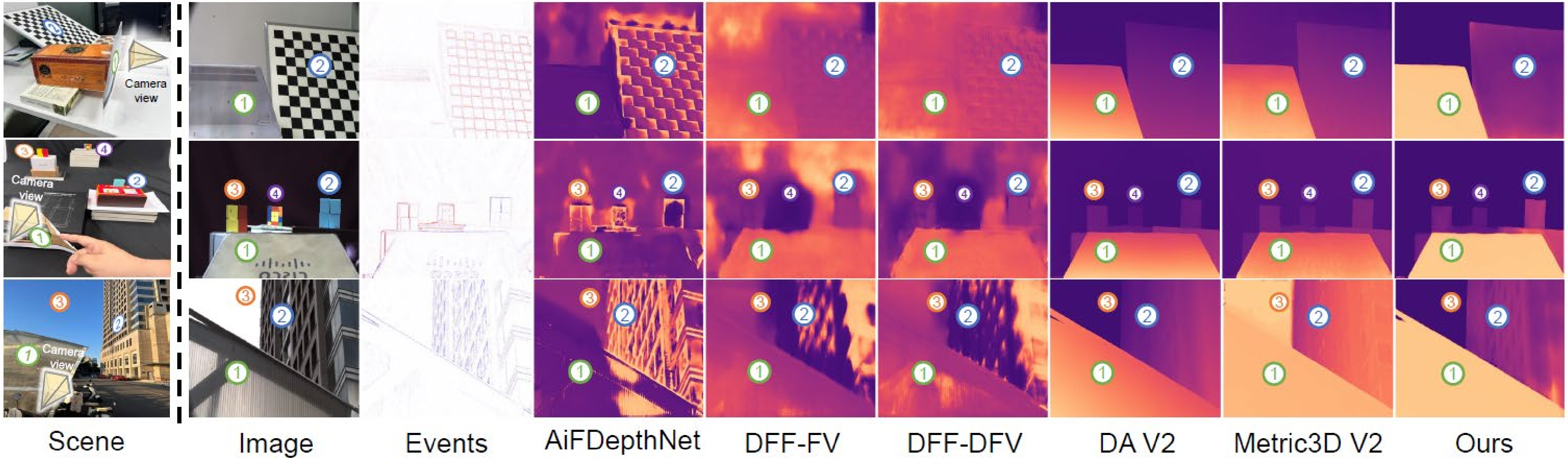
Method	Type	Blender-Syn						Sintel-Dr. Bokeh					
		RMSE(↓)	RMSE log(↓)	log10(↓)	δ_1(↑)	δ_2(↑)	δ_3(↑)	RMSE(↓)	RMSE log(↓)	log10(↓)	δ_1(↑)	δ_2(↑)	δ_3(↑)
DefocusNet	DFF	0.425	0.783	0.292	0.135	0.391	0.608	0.518	1.504	0.585	0.123	0.204	0.275
DFF-FV	DFF	0.325	0.661	0.162	0.669	0.732	0.775	0.267	0.982	0.343	0.177	0.364	0.518
DFF-DFV	DFF	0.369	0.710	0.196	0.651	0.681	0.707	0.270	1.038	0.366	0.192	0.332	0.495
DDFS	DFF	0.495	1.120	0.377	0.287	0.361	0.448	0.706	1.852	0.726	0.203	0.231	0.251
HybridDepth	DFF	0.622	1.461	0.570	0.050	0.127	0.227	0.442	1.391	0.551	0.110	0.186	0.262
DA V2	Mono	0.725	1.913	0.783	0.018	0.057	0.108	0.337	1.048	0.396	0.242	0.391	0.461
Metric3D V2	Mono	0.294	0.535	0.207	0.343	0.625	0.777	0.322	1.174	0.469	0.262	0.369	0.466
DAC	Mono	0.652	1.488	0.594	0.075	0.142	0.198	0.515	1.474	0.581	0.144	0.228	0.285
Ours	DFF	**0.148**	**0.282**	**0.081**	**0.697**	**0.878**	**0.944**	**0.233**	**0.685**	**0.253**	**0.333**	**0.466**	**0.560**



Results: Zero-shot Experiments

4DLFD-Semi-Real

Method	Type	RMSE(↓)	RMSE log(↓)	log10(↓)	δ_1(↑)	δ_2(↑)	δ_3(↑)
DFF-FV	DFF	1.979	0.198	0.070	0.680	0.888	0.949
DFF-DFV	DFF	1.943	0.186	0.064	0.711	0.902	0.953
DDFS	DFF	1.680	0.167	0.060	0.772	0.918	0.956
HybridDepth	DFF	1.676	0.137	0.051	0.827	0.949	0.957
DA V2	Mono	2.997	0.227	0.088	0.561	0.897	**0.958**
Metric3D V2	Mono	2.972	0.221	0.085	0.594	0.892	0.953
DAC	Mono	2.915	0.229	0.087	0.590	0.889	0.953
Ours	DFF	**1.549**	**0.128**	**0.047**	**0.832**	**0.957**	**0.958**



北京大学计算机学院
School of Computer Science

Camera Intelligence
A Computational Photography Lab @ PKU
http://camera.pku.edu.cn

NEURAL INFORMATION
PROCESSING SYSTEMS

Thanks for watching!

Lab page

https://camera.pku.edu.cn

Github page

https://github.com/liboyu02/EDFV