JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Learning to Fuse Monocular and Multi-view Cues for Multi-frame Depth Estimation in Dynamic Scenes

Rui Li[1], Dong Gong[2], Wei Yin[3], Hao Chen[4], Yu Zhu[1], Kaixuan Wang[3], Xiaozhi Chen[3], Jinqiu Sun[1], Yanning Zhang[1]

[1]Northwestern Polytechnical University, [2]The University of New South Wales, [3]DJI, [4]Zhejiang University

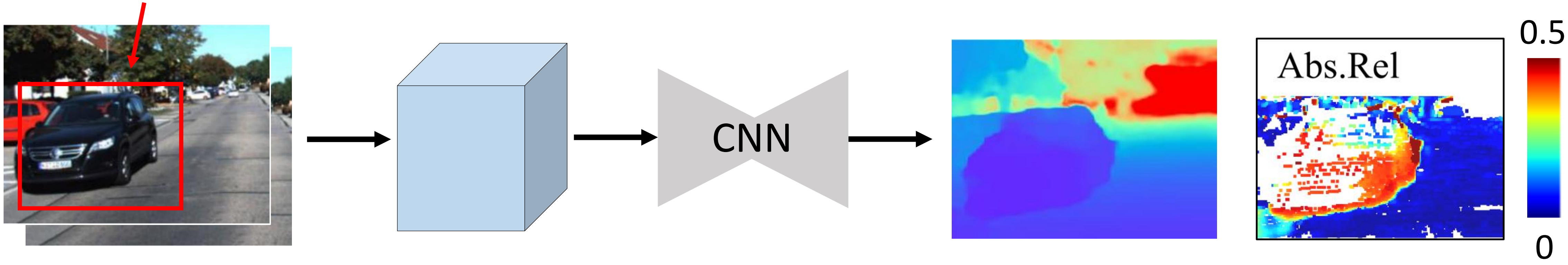
Project page: https://ruili3.github.io/dymultidepth/index.html

Github： https://github.com/ruili3/dynamic-multiframe-depth

THU-PM-089

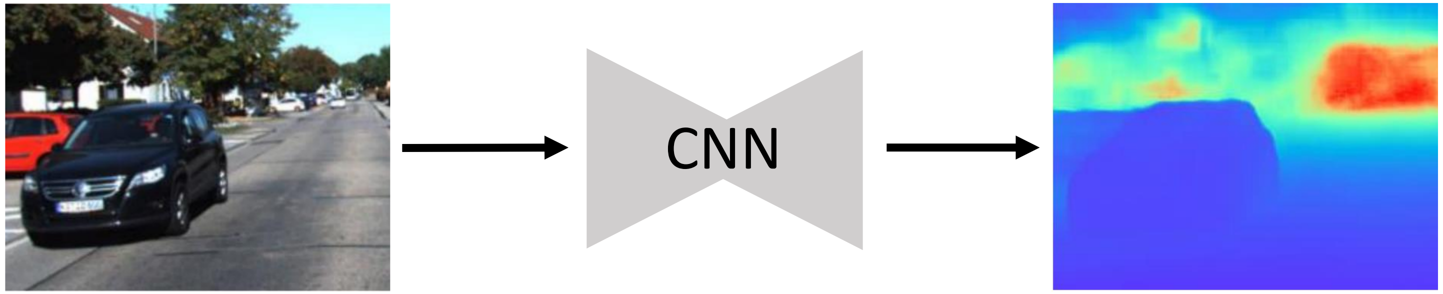
Multi-frame depth estimation

Higher general accuracy by leveraging multi-view consistency



Monocular depth estimation

Infer depth directly from a single image, not affected by dynamic issues.



Previous works

Segment dynamic areas, and *supplement* the multi-frame cues with monocular cues.

Limitations:

- Uncontrolled segmentation quality;
- Additional segmentation computation;
- Dynamic performance limited by monocular depth.

[1] MonoRec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera. CVPR 2021.
[2] The temporal opportunist: Self-supervised multi-frame monocular depth. CVPR 2021.
[3] Disentangling Object Motion and Occlusion for Unsupervised Multi-frame Monocular Depth. ECCV 2022.

Our work

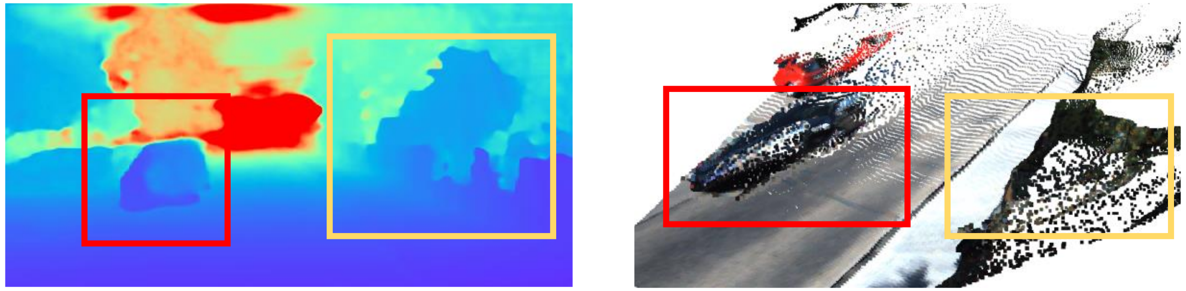
We propose a novel cross-cue fusion framework for dynamic depth estimation:

- Mask-free
- Obvious improvement on both cues (especially for mono. depth)

Insights

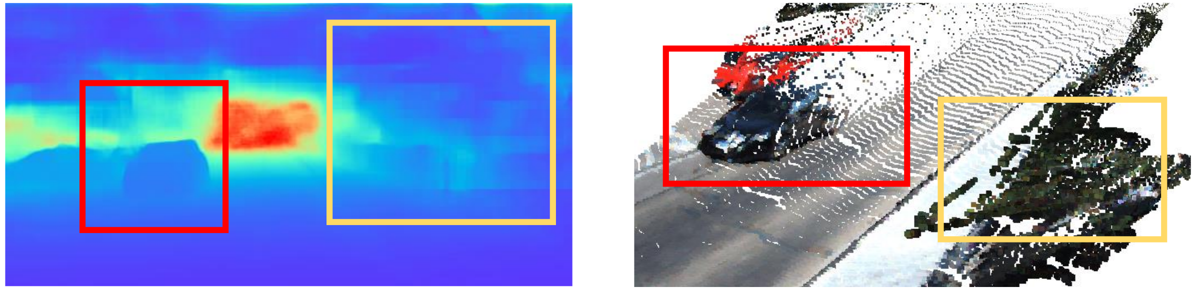
Two depth cues can potentially *benefit* each other due to their respective benefits on static and dynamic areas.

- Multi-frame depth



- Static
- Dynamic

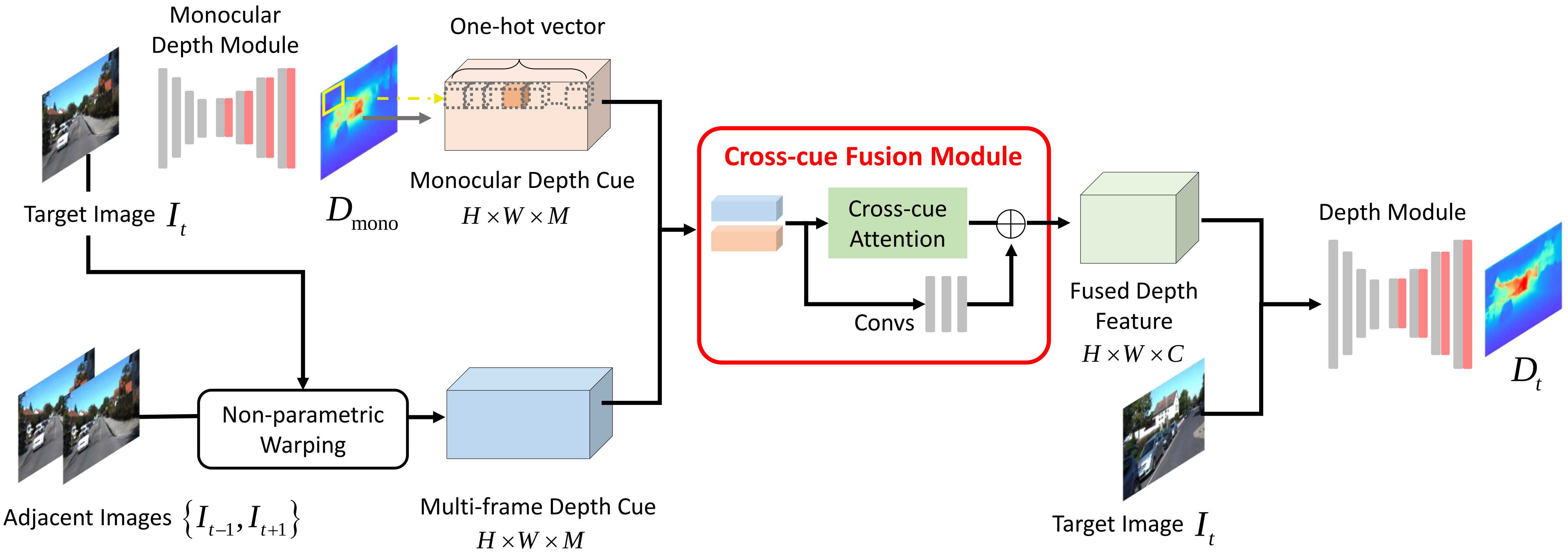
- Monocular depth



Depth Map Point Cloud

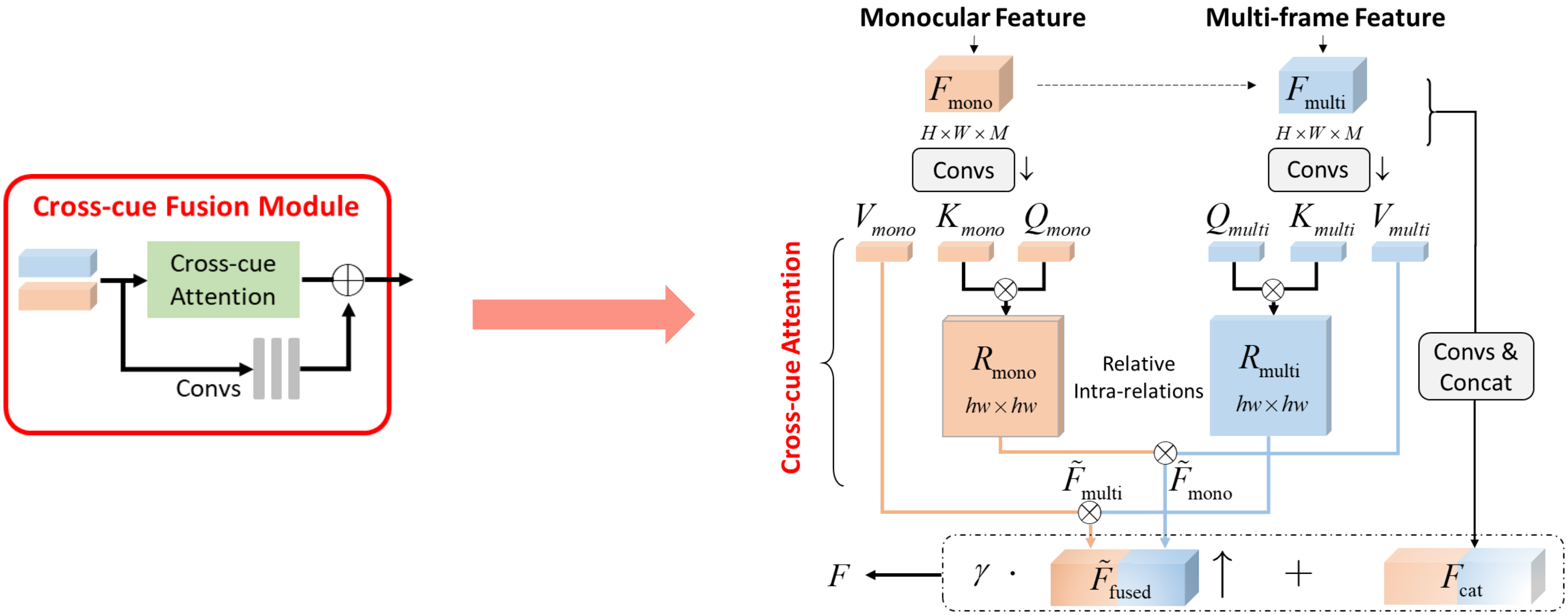
- Static
- Dynamic

Volume fusion with cross-cue attention



The cross-cue module

Enhance one depth feature with the learned intra-relations from another.



The cross-cue module

Taking multi-frame feature enhancement as an example:

Elements

$$Q_{\text{mono}} = f\left(F_{\text{mono}}, \theta^{Q}_{\text{mono}}\right),$$

$$K_{\text{mono}} = f\left(F_{\text{mono}}, \theta^{K}_{\text{mono}}\right),$$

$$V_{\text{multi}} = f\left(F_{\text{multi}}, \theta^{V}_{\text{multi}}\right)$$

Attn.

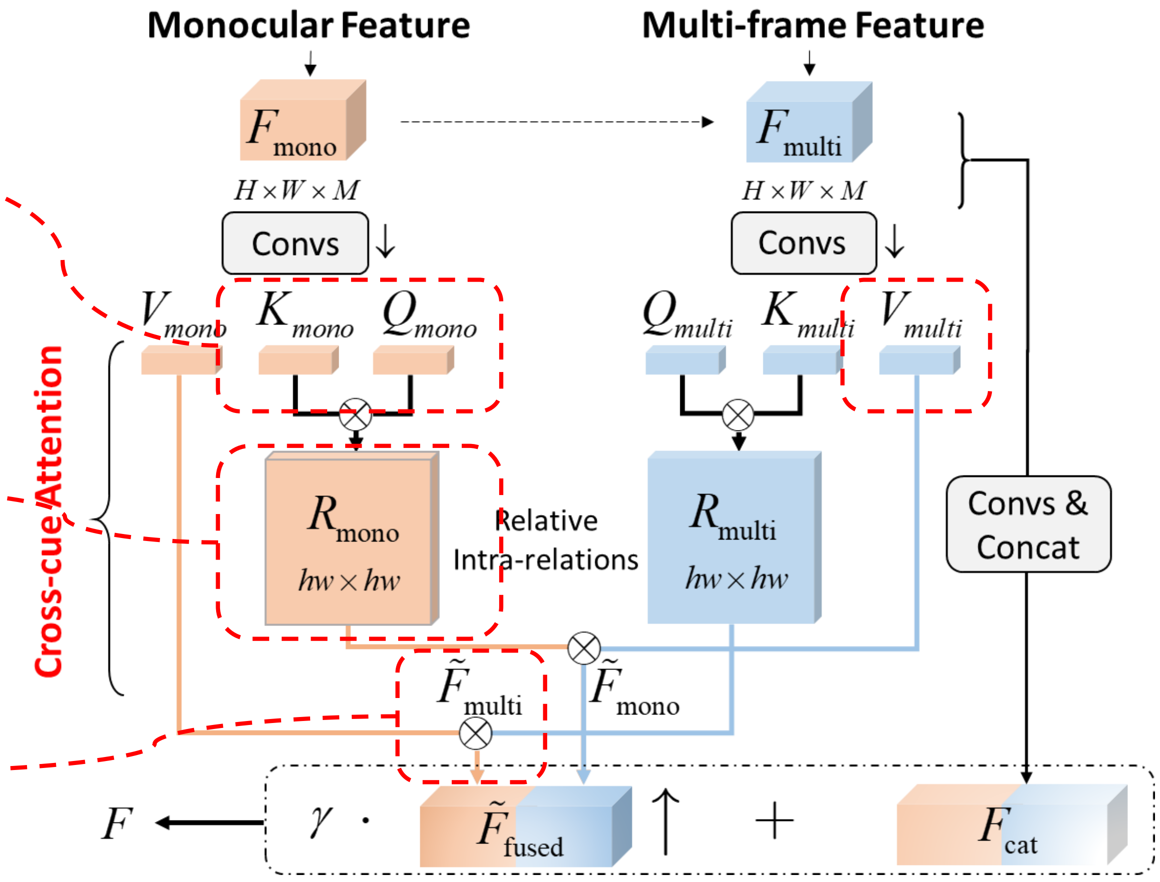
Monocular intra-relations

$$\boxed{R_{\text{mono}}} = \text{Softmax}\left(Q_{\text{mono}} \otimes K^{T}_{\text{mono}}\right)$$

Attending

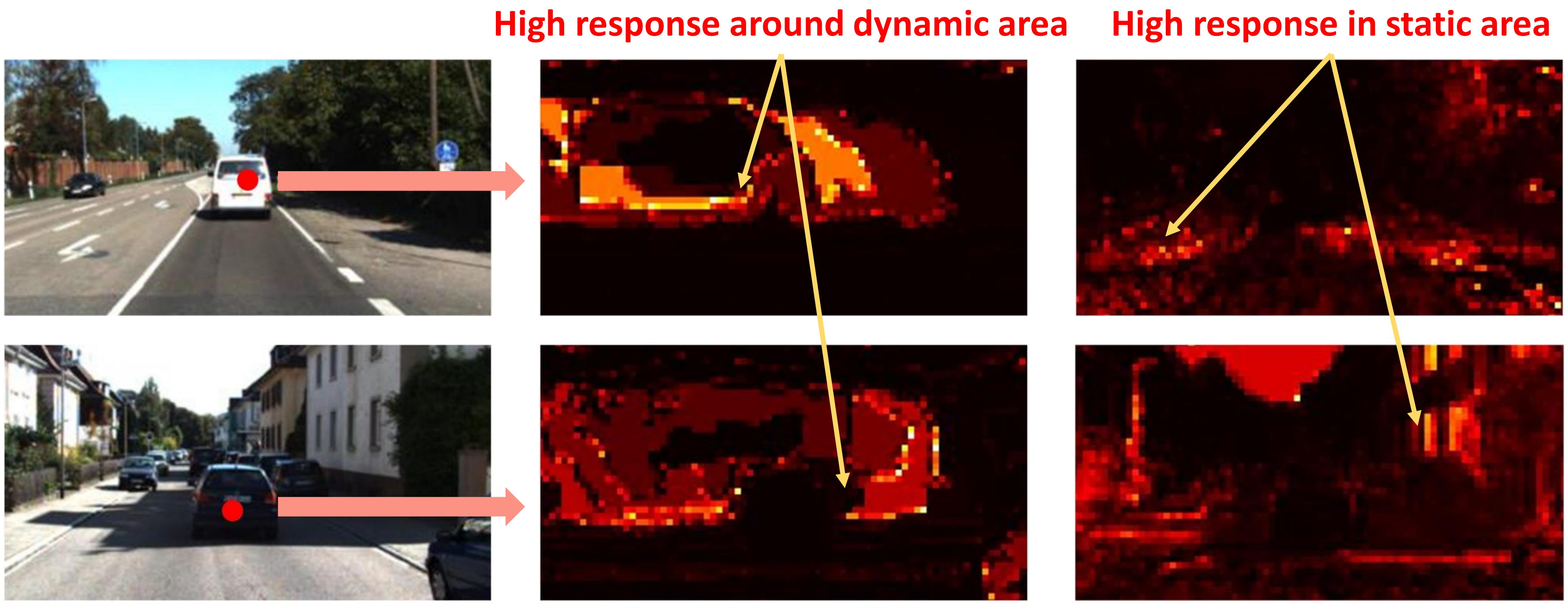
Enhanced multi-frame features

$$\boxed{\tilde{F}_{\text{multi}}} = R_{\text{mono}} \otimes V_{\text{multi}}$$



The cross-cue module

The effectiveness of intra-relations from each depth cue:



Input with dynamic point

Attn. map from **monocular** intra-relation R_{mono}

Attn. map from **multi-frame** intra-relation R_{multi}

Experiments

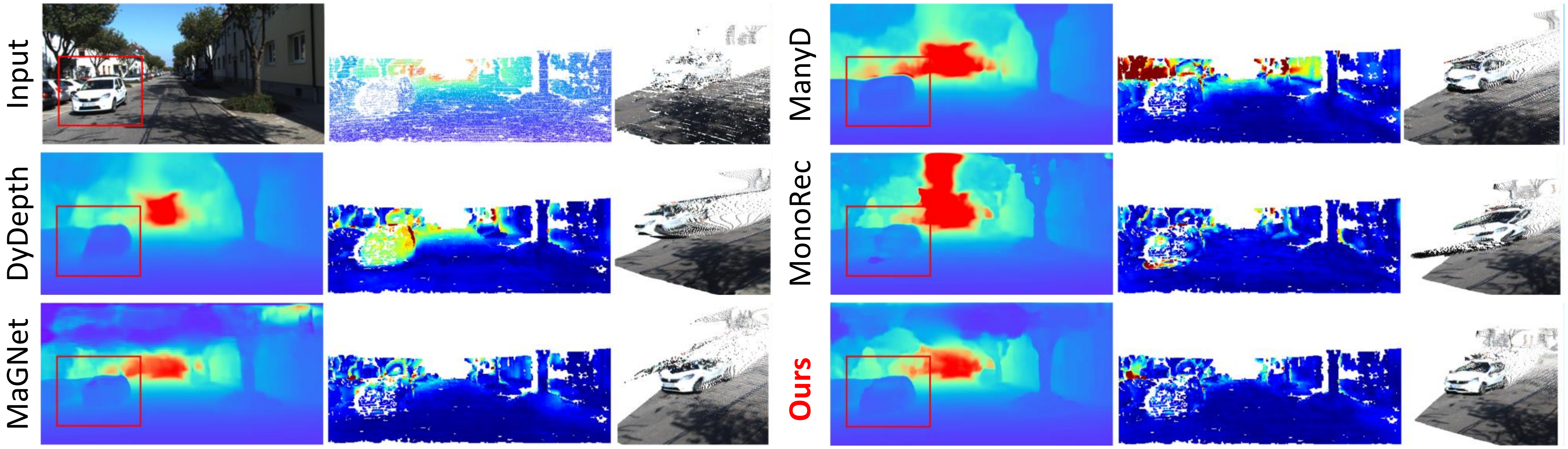
State-of-the-art overall & dynamic performance on KITTI

Eval	Method	Back.	Reso.	Sup.	Abs Rel	Sq Rel	RMSE	$RMSE_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Overall	Manydepth [36]	Res-18	MR	M	0.071	0.343	3.184	0.108	0.945	0.991	0.998
	DynamicDepth [9]	Res-18	MR	M	0.068	0.296	3.067	0.106	0.945	0.991	0.998
	MonoRec [37]	Res-18	MR	D*	0.050	0.290	2.266	0.082	0.972	0.991	0.996
	Ours	Res-18	MR	D	**0.043**	**0.151**	**2.113**	**0.073**	**0.975**	**0.996**	**0.999**
	MaGNet [1]	Effi-B5	MR	D	0.057	0.215	2.597	0.088	0.967	**0.996**	**0.999**
	Ours	Effi-B5	MR	D	0.046	0.155	2.112	0.076	0.973	**0.996**	**0.999**
	MaGNet [1]	Effi-B5	HR	D	0.043	0.135	2.047	0.082	0.981	**0.997**	**0.999**
	Ours	Effi-B5	HR	D	**0.039**	**0.103**	**1.718**	**0.067**	**0.981**	**0.997**	**0.999**
Dynamic	Manydepth [36]	Res-18	MR	M	0.222	3.390	7.921	0.237	0.676	0.902	0.964
	DynamicDepth [9]	Res-18	MR	M	0.208	2.757	7.362	0.227	0.682	0.911	0.971
	MonoRec [37]	Res-18	MR	D*	0.360	9.083	10.963	0.346	0.590	0.882	0.780
	Ours	Res-18	MR	D	0.118	0.835	4.297	0.146	0.871	0.975	0.990
	MaGNet [1]	Effi-B5	MR	D	0.141	1.219	4.877	0.168	0.830	0.955	0.986
	Ours	Effi-B5	MR	D	**0.111**	**0.768**	**4.117**	**0.135**	**0.881**	**0.980**	**0.994**
	MaGNet [1]	Effi-B5	HR	D	0.140	1.060	4.581	0.202	0.834	0.954	0.982
	Ours	Effi-B5	HR	D	**0.112**	**0.830**	**4.101**	**0.137**	**0.885**	**0.978**	**0.992**

Self-supervised Weakly-supervised Supervised

Experiments

Visualization of predicted depth map, error map and point cloud



Experiments

Good generalization results on DDAD

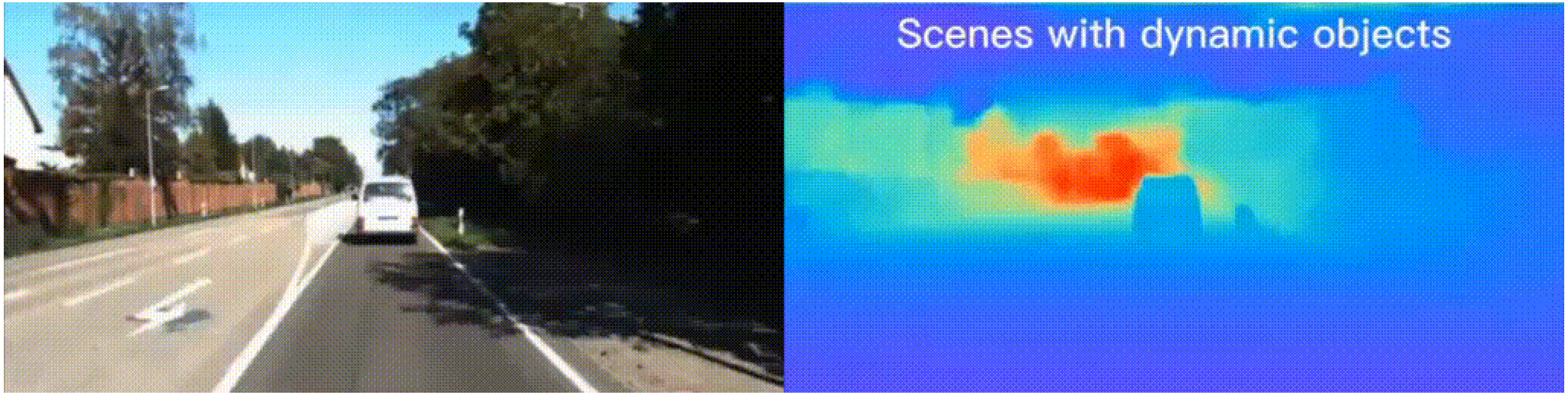
Eval	Method	Backbone	Abs Rel	Sq Rel	RMSE	$RMSE_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Overall	MonoRec [37]	Res-18	**0.158**	3.102	**7.553**	**0.227**	**0.854**	**0.931**	**0.961**
	MaGNet [1]	Effi-B5	0.208	<u>2.641</u>	10.739	0.382	0.620	0.878	0.942
	Ours	Res-18	**0.158**	**2.416**	<u>9.855</u>	<u>0.299</u>	<u>0.747</u>	<u>0.894</u>	<u>0.947</u>
Dynamic	MonoRec [37]	Res-18	0.544	16.703	16.116	0.482	0.460	0.667	0.798
	MaGNet [1]	Effi-B5	<u>0.266</u>	<u>3.982</u>	<u>11.715</u>	<u>0.398</u>	<u>0.462</u>	<u>0.815</u>	<u>0.917</u>
	Ours	Res-18	**0.234**	**3.611**	**11.007**	**0.331**	**0.576**	**0.835**	**0.921**

Experiments

Dynamic depth error reduction over the monocular depth branch.

Method	Mono. Err.	Final Err.	Err. Redu.
Manydepth [36]	0.212	0.222	-4.72%
Dynamicdepth [9]	0.214	0.208	2.83%
MaGNet [1]	0.153	0.141	7.84%
Ours - Res.18	0.149	0.118	**20.81%**
Ours - Res.50	0.145	0.116	**20.00%**

Thank you!



Project page: https://ruili3.github.io/dymultidepth/index.html
Github： https://github.com/ruili3/dynamic-multiframe-depth