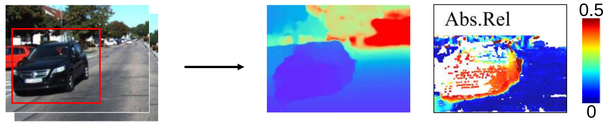
Learning to Fuse Monocular and Multi-view Cues for Multi-frame Depth Estimation in Dynamic Scenes

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

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Multi-frame Depth in Dynamic Areas

Problem Statement: Multi-frame depth estimation encounters severe corruption in dynamic areas, due to the violation of multi-view consistency.



Inputs with dynamic area

Depth map showing large dynamic error

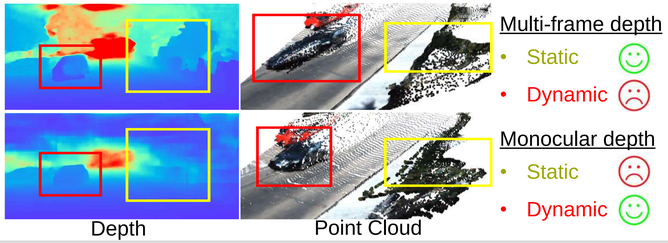
Existing Methods: Segment the dynamic areas, supplement the dynamic multi-frame cue with the monocular depth cue.

Limitations:

- Dynamic area segmentation is challenging, with additional computation overhead;
- The single dependency of monocular cues limits the dynamic depth performance.

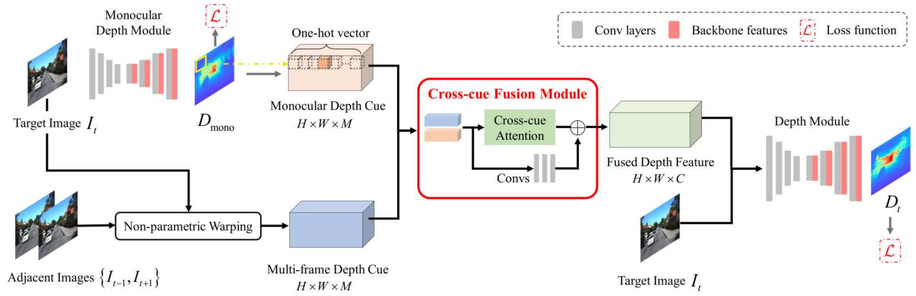
Mutual Benefits of Two Depth Cues

We aim to propagate the multi-frame static (**yellow** box) depth to the monocular cues and let monocular cues in dynamic areas (**red** box) enhance the multi-frame representations, yielding the final dynamic depth *excelling* each depth cue.

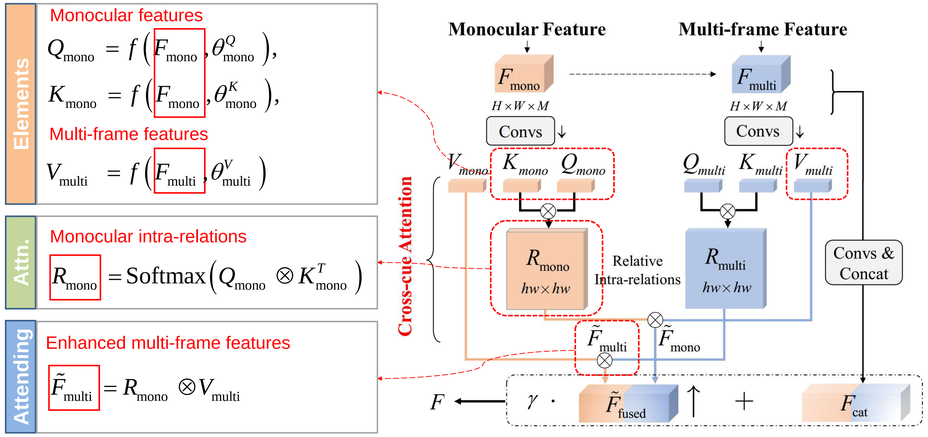


Enhancing Multi-frame & Monocular Cues with Cross-cue Fusion

Volume Fusion with Cross-cue Attention: We let each depth cue benefit the other by volume fusion with the Cross-cue Fusion (CCF) module, yielding a *mask-free* approach.



Cross-cue Attention: We generate *query*, *key* features from one depth cue to compute its relative intra-relations, then use it to enhance the *value* feature from the other depth cue. Take the multi-frame feature enhancing process as an example:

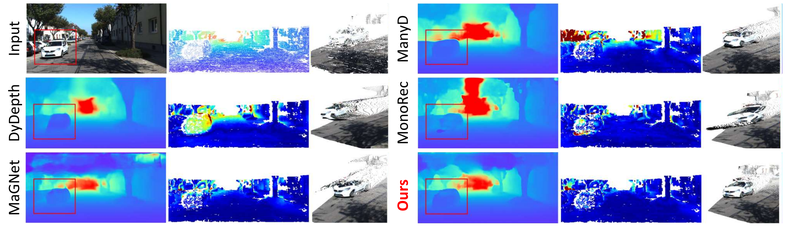


Experiments

KITTI: Evaluation of *overall* & *dynamic* depth errors.

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**

From left to right: depth map, error map, and reconstructed point cloud.



Dynamic error reduction over monocular branch:

Our method achieves significant dynamic 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%**

Find the code and models here!