Supplementary:BRNet: Exploring Comprehensive Features for Monocular Depth Estimation

Wencheng Han[1], Junbo Yin[2], Xiaogang Jin[3], Xiangdong Dai[4], and Jianbing Shen[1⋆],

[1] SKL-IOTSC, Computer and Information Science, University of Macau
[2] School of Computer Science, Beijing Institute of Technology
[3] State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, China
[4] Guangdong OPPO Mobile Telecommunications Corp., Ltd

1 Results of Different Supervision Types

We compare the results of three training supervision types, M (Monocular videos), S (Stereo image pairs) and MS (Monocular video and Stereo image pairs). As shown in Table 2 and Fig. 1, BRNet with MS training achieves the best result among the three types. The model trained on stereo image pairs shows better results in Abs Rel but worse results in other metrics than monocular video training.

2 Improved Ground Truth

The evaluation method introduced by Eigen [1] for KITTI uses the projected LIDAR points. However, it does not handle occlusions or moving objects, which are very common because the cars are usually travelling. To alleviate these problems, [10] introduced a set of high-quality depth maps for the KITTI dataset, which is made by five consecutive frames, and the moving objects are handled by stereo pairs. The improved ground truth contains 652 frames from the Eigen split, which is 93% of the total test frames (697). Following [3], we evaluate our methods on these frames with improved ground truth and compare them to several representative networks.

We employ the same error metrics from the standard evaluation and clip the predicted depth to 80m to match the Eigen evaluation. As shown in Table 1, our methods trained by all the three supervision types achieve significant improvements from our baseline work, outperforming all existing methods.

3 Effective of Post-Processing

Post-processing in depth-estimation introduced by [2] is a technique to improve test time results. Specifically, with post-process, the model will run each test

⋆ Corresponding author: *Jianbing Shen*, email: shenjianbingcg@email.com.

Method	Resolution	Train	lower is better				higher is better		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta^2 < 1.25$	$\delta^3 < 1.25$
SfMLearner [9]	416 × 128	M	0.176	1.532	6.129	0.244	0.758	0.921	0.971
VidDepth [7]	416 × 128	M	0.134	0.983	5.501	0.203	0.827	0.944	0.981
GeoNet [12]	416 × 128	M	0.132	0.994	5.240	0.193	0.883	0.953	0.985
DDVO [11]	416 × 128	M	0.126	0.866	4.932	0.185	0.851	0.958	0.986
EPC++ [5]	640 × 192	M	0.120	0.789	4.755	0.177	0.856	0.961	0.987
Monodepth2 [3]	640 × 192	M	0.090	0.545	3.942	0.137	0.914	0.983	0.995
BRNet	640 × 192	M	**0.080**	**0.409**	**3.613**	**0.124**	**0.928**	**0.987**	**0.997**
Monodepth [2]	512 × 256	S	0.109	0.811	4.568	0.166	0.877	0.967	0.988
3Net(VGG) [9]	640 × 192	S	0.119	0.920	4.824	0.182	0.856	0.957	0.985
3Net(ResNet50) [9]	640 × 192	S	0.102	0.675	4.293	0.159	0.881	0.969	0.991
SuperDepth+pp [8]	416 × 128	S	0.090	0.542	3.967	0.144	0.901	0.976	0.993
Monodepth2 [3]	640 × 192	S	0.085	0.537	3.868	0.139	0.912	0.979	0.993
BRNet	640 × 192	S	**0.078**	**0.448**	**3.547**	**0.125**	**0.928**	**0.985**	**0.995**
EPC++ [5]	640 × 192	MS	0.123	0.754	4.453	0.172	0.863	0.964	0.989
Monodepth2 [3]	640 × 192	MS	0.080	0.466	3.681	0.127	0.926	0.985	0.995
BRNet	640 × 192	MS	**0.078**	**0.393**	**3.400**	**0.120**	**0.928**	**0.988**	**0.997**

Table 1. Comparison on KITTI improved ground truth. Comparison to other networks on 93% KITTI 2015 Eigen split [1] and improve ground truth from [10].

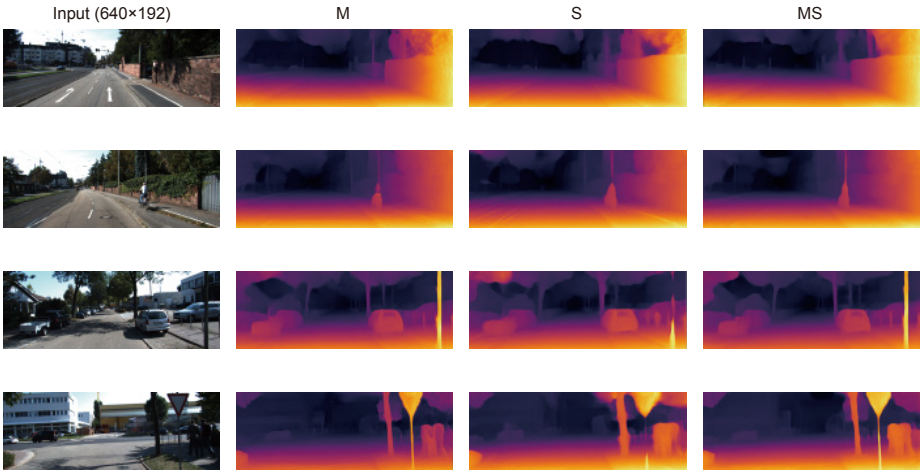

Fig. 1. Comparison of BRNet with different training methods.

image twice, once unflipped and then flipped. Then the flipped results are flipped back, and the two results are averaged as the final results. It has been proved to bring significant improvements in accuracy [3,9,2]. Followed by monodepth2 [3], we apply post-process on our model with three different training settings and two resolutions.

As shown in Table 2, with post-process, BRNet achieves obvious gains on all supervision types and resolutions. Especially, BRNet with MS training and large input(1024×320) achieves 0.095 and 4.298 in terms of Abs Rel and RMSE.

4 Inputs Resolutions of BRNet

As shown in previous works [3,6,4], higher input resolution can bring performance improvements. Our BRNet can extract more detailed information by the clearer

Method	Resolution	PostProcess	Train	lower is better				higher is better		
				Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta^2 < 1.25$	$\delta^3 < 1.25$
Monodepth2 [3]	640 × 192		M	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Monodepth2 [3]	640 × 192	✓	M	0.112	0.851	4.754	0.190	0.881	0.960	0.981
BRNet	640 × 192		M	0.105	0.698	4.462	0.179	0.890	**0.965**	**0.984**
BRNet	640 × 192	✓	M	**0.104**	**0.681**	**4.419**	**0.178**	**0.891**	**0.965**	**0.984**
Monodepth2 [3]	640 × 192		S	0.109	0.873	4.960	0.209	0.864	0.948	0.975
Monodepth2 [3]	640 × 192	✓	S	0.108	0.842	4.891	0.207	0.866	0.949	0.976
BRNet	640 × 192		S	0.103	0.792	4.716	0.197	0.876	0.954	**0.978**
BRNet	640 × 192	✓	S	**0.102**	**0.774**	**4.679**	**0.196**	**0.879**	**0.955**	**0.978**
Monodepth2 [3]	640 × 192		MS	0.106	0.818	4.750	0.196	0.874	0.957	0.979
Monodepth2 [3]	640 × 192	✓	MS	0.104	0.786	4.687	0.194	0.876	0.958	0.980
BRNet	640 × 192		MS	0.099	0.685	4.453	0.183	0.885	0.962	**0.983**
BRNet	640 × 192	✓	MS	**0.098**	**0.671**	**4.418**	**0.178**	**0.886**	**0.963**	**0.983**
Monodepth2 [3]	1024 × 320		M	0.115	0.882	4.701	0.190	0.879	0.961	0.982
Monodepth2 [3]	1024 × 320	✓	M	0.112	0.838	4.607	0.187	0.883	0.962	0.982
BRNet	1024 × 320		M	0.103	0.684	4.385	**0.175**	0.889	0.965	**0.985**
BRNet	1024 × 320	✓	M	**0.102**	**0.683**	**4.336**	0.175	**0.894**	**0.966**	**0.985**
Monodepth2 [3]	1024 × 320		S	0.107	0.849	4.764	0.201	0.874	0.953	0.977
Monodepth2 [3]	1024 × 320	✓	S	0.105	0.822	4.692	0.199	0.876	0.954	0.977
BRNet	1024 × 320		S	0.097	0.729	4.510	0.191	0.886	**0.958**	**0.979**
BRNet	1024 × 320	✓	S	**0.096**	**0.710**	**4.459**	**0.190**	**0.887**	**0.958**	**0.979**
Monodepth2 [3]	1024 × 320		MS	0.106	0.806	4.630	0.193	0.876	0.958	0.980
Monodepth2 [3]	1024 × 320	✓	MS	0.104	0.775	4.562	0.191	0.878	0.959	0.981
BRNet	1024 × 320		MS	0.097	0.677	4.378	0.179	0.888	**0.965**	**0.984**
BRNet	1024 × 320	✓	MS	**0.095**	**0.653**	**4.298**	**0.181**	**0.889**	0.964	0.983

Table 2. Results of BRNet on KITTI Eigen split with different supervision types and post process. M means monocular videos only and S means stereo image pairs, and MS means both. The best two results are shown in bold and underlined, respectively.

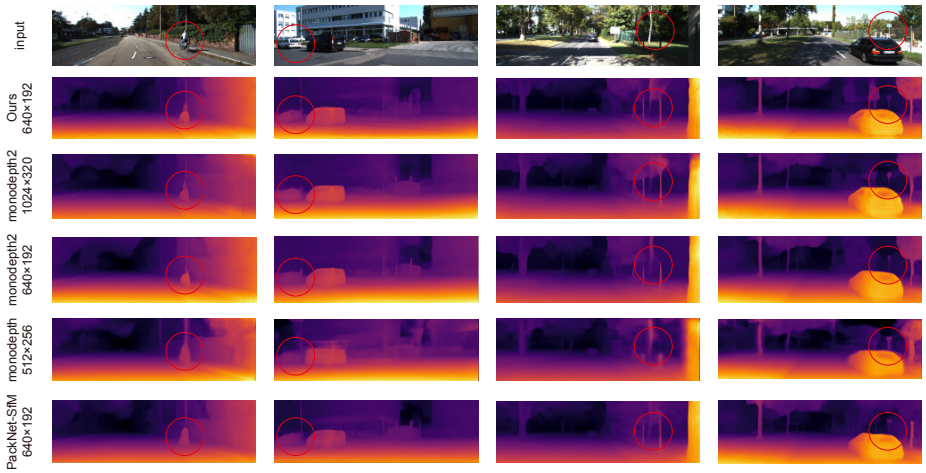

Fig. 2. Additional qualitative results on the KITTI Eigen split test set.

images and smaller receptive fields when taking large inputs. At the same time, with the global branch, BRNet will not lose critical global information. Thus, BRNet achieves better results when taking large inputs, as shown in Table 2.

We also show the qualitative results of different resolutions in Fig. 4. Results of large inputs show clear outlines of objects than small ones on all three training supervision types.

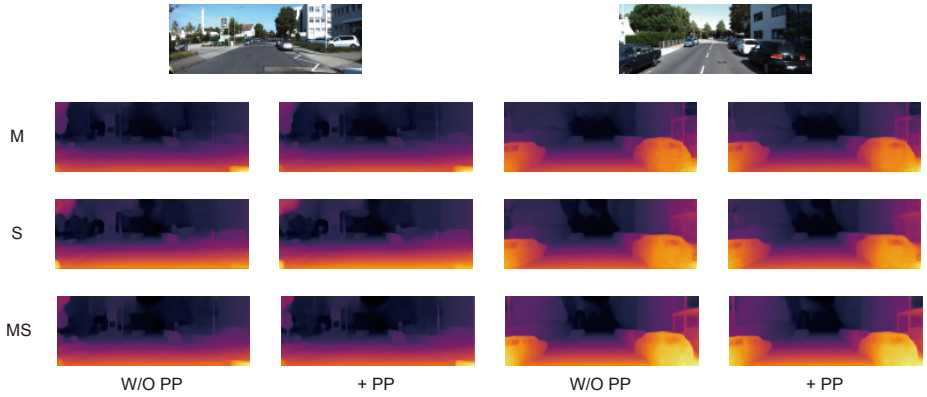

Fig. 3. Results of BRNet with or without post process

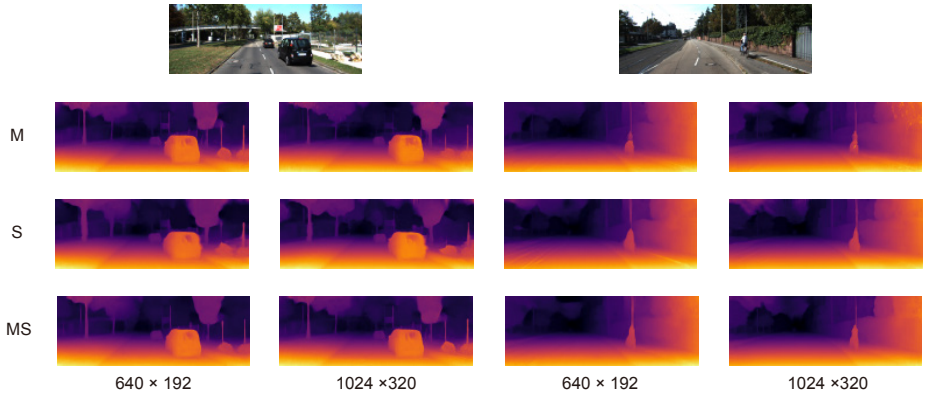

Fig. 4. Qualitative results BRNet with different resolutions.

5 Additional Qualitative Results

To clearly compare BRNet and existing networks, we present more qualitative results in Fig. 2. We select monodepth2 [3], monodepth [2] and PackNet-SfM[4] as our competitor, and for monodepth2 we extract results from both large and small input sizes.

As shown in the figure, our method gives the clearest prediction among all methods. We mark the most obvious region by the red circles. For monodepth2, results from higher resolution perform better than small resolution, especially for objects far from the camera, while BRNet taking small inputs can give even clearer prediction from these small objects.

References

1. David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015.
2. Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.
3. Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019.
4. Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2020.
5. Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *PAMI*, 42(10):2624–2641, 2019.
6. Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: high resolution self-supervised monocular depth estimation. *CoRR abs/2012.07356*, 2020.
7. Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 2018.
8. Sudeep Pillai, Rareş Ambruş, and Adrien Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. In *ICRA*, 2019.
9. Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *3DV*, 2018.
10. Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *3DV*, 2017.
11. Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, 2018.
12. Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018.