Supplementary Material for "RA-Depth: Resolution Adaptive Self-Supervised Monocular Depth Estimation"

Mu He, Le Hui, Yikai Bian, Jian Ren, Jin Xie*, and Jian Yang*

Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education
Jiangsu Key Lab of Image and Video Understanding for Social Security
PCA Lab, School of Computer Science and Engineering
Nanjing University of Science and Technology, China
{muhe,le.hui,yikai.bian,renjian,csjxie,csjyang}@njust.edu.cn

A Overview

In this documentation, we provide additional technical details, quantitative results, and qualitative results for our method RA-Depth. In Sec. B, we first give the details of the monocular depth estimation network. Then, we provide additional evaluation for RA-Depth in Sec. C. Finally, we show more visualization results on the KITTI dataset [2] and internet photos in Sec. D.

B Network Details

We use our proposed Dual HRNet as the monocular depth estimation network. Dual HRNet uses HRNet18 [9] as the encoder named as HREncoder and the proposed HRDecoder as the decoder. The implementation details of HRDecoder are shown in Fig. 1.

C Additional Evaluation

Ranges of Scale Factors s^L and s^H. As shown in Table 1, we report how the ranges of scale factors s^L and s^H in the arbitrary-scale data augmentation component affect the results. Specifically, Range2 represents the scale range used in the main paper, while Range1 and Range3 represent smaller and larger scale variation ranges, respectively. All models are trained at the resolution of 640×192 on the KITTI dataset [2] using the Eigen split [1]. Experiments show that in terms of varying resolutions, the setting of Range2 achieves the best results for depth estimation in most cases.

* Corresponding authors.

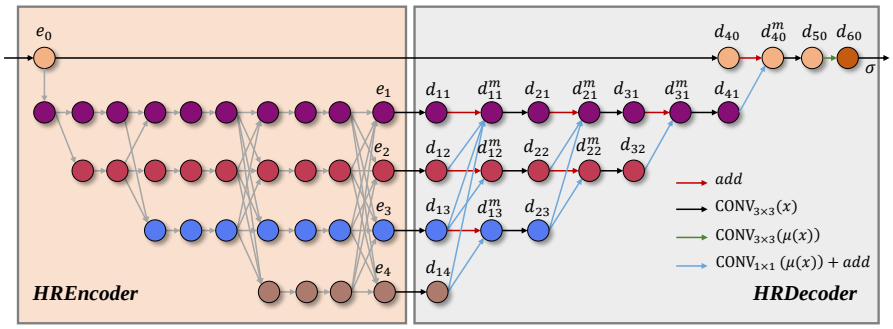

Fig. 1. Network Details for Dual HRNet. *add*: a summation operator. $\text{CONV}_{3\times3}$: a 3×3 convolution layer. $\text{CONV}_{1\times1}$: an 1×1 convolution layer. $\mu(\cdot)$: an upsampling operator. $\sigma(\cdot)$: a sigmoid activation function.

Monodepth2 With Arbitrary-Scale Data Augmentation. As shown in Table 2, we use the arbitrary-scale data augmentation (dubbed AS-Aug) on Monodepth2 [3] for depth estimation. Experimental results show that the proposed arbitrary-scale data augmentation can boost the performance of Monnodepth2.

Comparisons With Other Individual Models. As shown in Table 3, we compare our single model RA-Depth trained at the fixed resolution of 640×192 on the KITTI dataset [2] with other individual models [3,5,10] trained separately at each test resolution. It can be observed that the performance of our RA-Depth significantly outperforms other state-of-the-art methods. Although our model is trained only once at a fixed resolution, our model can achieve the best results across different test resolutions.

C.1 Improved Ground Truth

[8] has introduced a set of high-quality depth maps for the KITTI dataset [2], resulting in 652 improved ground-truth depth maps for testing. These 652 improved ground-truth depth maps are provided for 652 (or 97%) of the 697 test frames contained in the Eigen test split [1]. As shown in Table 4, we evaluate our RA-Depth on these 652 improved ground-truth depth maps. Note that for a fair comparison, we use the same evaluation criteria and metrics as Monodepth2 [3]. It can be observed that RA-Depth still significantly outperforms existing state-of-the-art self-supervised approaches [11,4,3,5,10].

D More Visualization Results

Visualization Results of Public Datasets. As shown in Fig. 2, we show more visualization results on the KITTI dataset [2] using the Eigen test split [1].

Table 1. Experiments for the ranges of scale factors s^L and s^H. Range1: $s^L \in [0.8, 0.9]$, $s^H \in [1.1, 1.5]$. Range2: $s^L \in [0.7, 0.9]$, $s^H \in [1.1, 2.0]$. Range3: $s^L \in [0.5, 0.9]$, $s^H \in [1.1, 3.0]$. All models are trained at the resolution of 640×192 and then tested at five different resolutions including including 416×128, 512×160, 640×192, 832×256, and 1024×320. The best results are in **bold** for each test resolution.

Range	Test Resolution	Error Metric ↓				Accuracy Metric ↑		
		AbsRel	SqRel	RMSE	RMSElog	δ<1.25	δ<1.25^2	δ<1.25^3
Range1	416×128	0.113	0.732	**4.632**	**0.185**	**0.876**	**0.963**	**0.984**
Range2	416×128	**0.111**	**0.723**	4.768	0.187	0.874	0.961	**0.984**
Range3	416×128	0.119	0.758	4.969	0.193	0.861	0.958	0.983
Range1	512×160	0.103	0.694	4.404	0.177	0.892	0.965	0.984
Range2	512×160	**0.101**	**0.658**	**4.373**	**0.175**	**0.895**	**0.967**	**0.985**
Range3	512×160	0.107	0.668	4.500	0.180	0.885	0.965	**0.985**
Range1	640×192	0.099	0.662	4.302	0.174	0.898	0.966	0.984
Range2	640×192	**0.096**	**0.632**	**4.216**	**0.171**	**0.903**	**0.968**	**0.985**
Range3	640×192	0.100	0.641	4.281	0.174	0.896	0.967	**0.985**
Range1	832×256	0.102	0.641	4.236	0.176	0.895	0.967	0.984
Range2	832×256	**0.095**	0.613	**4.106**	**0.170**	**0.906**	**0.969**	**0.985**
Range3	832×256	0.098	**0.610**	4.135	0.171	0.901	0.968	**0.985**
Range1	1024×320	0.112	0.669	4.310	0.185	0.878	0.964	0.984
Range2	1024×320	**0.097**	**0.608**	**4.131**	**0.174**	**0.901**	**0.969**	**0.985**
Range3	1024×320	0.103	0.611	4.126	0.175	0.898	**0.969**	**0.985**

Table 2. We perform experiments on Monodepth2 [3] by using our proposed arbitrary-scale augmentation.

Method	Train Resolution	Test Resolution	AbsRel	SqRel	RMSE	RMSElog
Monodepth2	640×192	416×128	0.184	1.365	6.146	0.268
Monodepth2 + AS-Aug	640×192	416×128	**0.116**	**0.900**	**4.902**	**0.193**
Monodepth2	640×192	1024×320	0.193	1.335	6.058	0.271
Monodepth2 + AS-Aug	640×192	1024×320	**0.106**	**0.853**	**4.607**	**0.182**

In addition, Fig. 3 shows the visualization results of depth estimation on the Make3D and NYU-V2 datasets.

Visualization Results of Internet Photos. As shown in Fig. 4, we use the RA-Depth model trained on the KITTI dataset to predict the depth maps for these images from the 'Wind Walk Travel Videos' YouTube channel[1]. These images are captured with a monocular hand-held camera and are quite different from the car-mounted videos of the KITTI dataset [2]. It can be observed that our RA-Depth can predict higher quality depth maps than existing methods [3,5,10] on these internet photos.

[1] https://www.youtube.com/channel/UCPur06mx78RtwgHJzxpu2ew

Table 3. Comparisons with other individual models on the KITTI dataset using the Eigen split [1]. Our model RA-Depth is trained at the resolution of 640×192 and then test at five different test resolutions including 416×128, 512×160, 640×192, 832×256, and 1024×320. Existing superior methods [3,5,10] train an individual model for each test resolution. In each category of test resolution, the best results are in **bold** and the second are underlined.

Method	Train Resolution	Test Resolution	Error Metric ↓				Accuracy metric ↑		
			AbsRel	SqRel	RMSE	RMSElog	δ<1.25	δ<1.25^2	δ<1.25^3
Monodepth2 [3]	416×128	416×128	0.126	1.026	5.143	0.203	0.855	0.954	0.979
HR-Depth [5]	416×128	416×128	0.119	0.920	4.924	0.196	0.866	0.957	0.980
DIFFNet [10]	416×128	416×128	0.114	0.884	4.829	0.191	0.874	0.958	0.981
RA-Depth	640×192	416×128	**0.111**	**0.723**	**4.768**	**0.187**	**0.874**	**0.961**	**0.984**
Monodepth2 [3]	512×160	512×160	0.119	0.946	4.969	0.197	0.868	0.957	0.980
HR-Depth [5]	512×160	512×160	0.113	0.852	4.766	0.190	0.875	0.959	0.982
DIFFNet [10]	512×160	512×160	0.108	0.778	4.605	0.184	0.885	0.963	0.983
RA-Depth	640×192	512×160	**0.101**	**0.658**	**4.373**	**0.175**	**0.895**	**0.967**	**0.985**
Monodepth2 [3]	640×192	640×192	0.115	0.903	4.863	0.193	0.877	0.959	0.981
HR-Depth [5]	640×192	640×192	0.109	0.792	4.632	0.185	0.884	0.962	0.983
DIFFNet [10]	640×192	640×192	0.102	0.764	4.483	0.180	0.896	0.965	0.983
RA-Depth	640×192	640×192	**0.096**	**0.632**	**4.216**	**0.171**	**0.903**	**0.968**	**0.985**
Monodepth2 [3]	832×256	832×256	0.111	0.859	4.689	0.187	0.885	0.962	0.982
HR-Depth [5]	832×256	832×256	0.108	0.781	4.563	0.183	0.887	0.964	0.983
DIFFNet [10]	832×256	832×256	0.101	0.761	4.441	0.177	0.899	0.966	0.983
RA-Depth	640×192	832×256	**0.095**	**0.613**	**4.106**	**0.170**	**0.906**	**0.969**	**0.985**
Monodepth2 [3]	1024×320	1024×320	0.108	0.863	4.647	0.186	0.892	0.963	0.982
HR-Depth [5]	1024×320	1024×320	0.106	0.755	4.472	0.181	0.892	0.966	0.984
DIFFNet [10]	1024×320	1024×320	**0.097**	0.722	4.345	**0.174**	**0.907**	0.967	0.984
RA-Depth	640×192	1024×320	**0.097**	**0.608**	**4.131**	**0.174**	0.901	**0.968**	**0.985**

Table 4. Results on improved ground truth depth. Comparison to existing self-supervised approaches on the KITTI dataset [2] using 93% of the Eigen split [1] and the improved ground truth from [8].

Method	Resolution	Error Metric ↓				Accuracy metric ↑		
		AbsRel	SqRel	RMSE	RMSElog	δ<1.25	δ<1.25^2	δ<1.25^3
SfMLearner [11]	640×192	0.176	1.532	6.129	0.244	0.758	0.921	0.971
EPC++ [4]	832×256	0.120	0.789	4.755	0.177	0.856	0.961	0.987
Monodepth2 [3]	640×192	0.090	0.545	3.942	0.137	0.914	0.983	0.995
HR-Depth [5]	640×192	0.079	0.421	3.603	0.123	0.928	0.987	0.997
DIFFNet [10]	640×192	0.076	0.414	3.492	0.119	0.936	0.988	0.996
RA-Depth	640×192	**0.074**	**0.362**	**3.345**	**0.113**	**0.940**	**0.990**	**0.997**

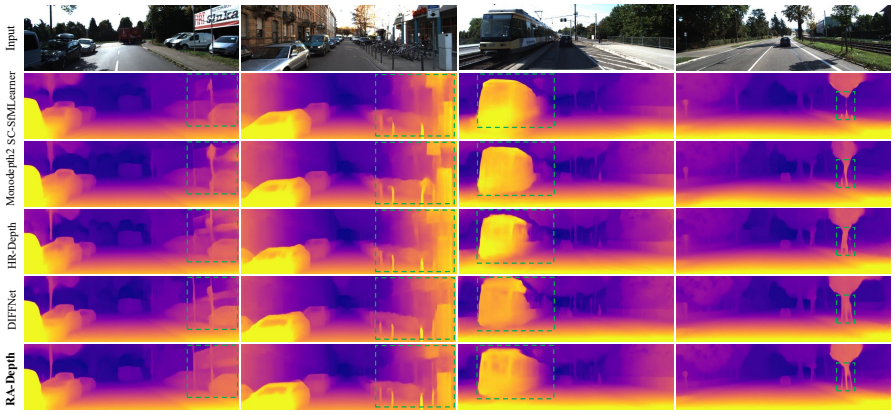

Fig. 2. Additional qualitative results on the KITTI dataset using Eigen split.

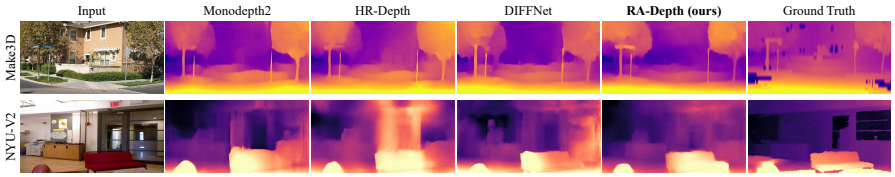

Fig. 3. Visualization results on the Make3D [6] and NYU-V2 [7] datasets. All models are trained on the KITTI dataset [2].

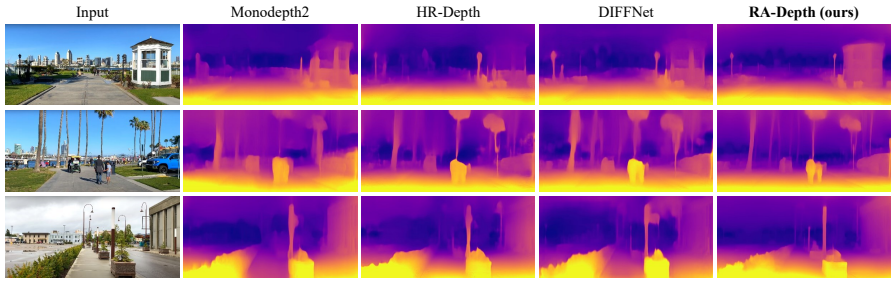

Fig. 4. Additional qualitative results on the internet photos captured with a monocular hand-held camera.

References

1. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. arXiv preprint arXiv:1406.2283 (2014)
2. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research (2013)
3. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: ICCV (2019)
4. Luo, C., Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R., Yuille, A.: Every pixel counts++: Joint learning of geometry and motion with 3D holistic understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
5. Lyu, X., Liu, L., Wang, M., Kong, X., Liu, L., Liu, Y., Chen, X., Yuan, Y.: Hr-depth: High resolution self-supervised monocular depth estimation. AAAI (2020)
6. Saxena, A., Sun, M., Ng, A.Y.: Make3D: Learning 3D scene structure from a single still image. IEEE transactions on pattern analysis and machine intelligence (2008)
7. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
8. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant cnns. In: 3DV (2017)
9. Wang, J., Sun, K., Cheng, T.: Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
10. Zhou, H., Greenwood, D., Taylor, S.: Self-supervised monocular depth estimation with internal feature fusion. In: BMVC (2021)
11. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017)