Supplementary Material for RM-Depth: Unsupervised Learning of Recurrent Monocular Depth in Dynamic Scenes*

Tak-Wai Hui
H-1 Research
eetwhui@gmail.com

1. Odometry Evaluation

The pose network of RM-Depth [6] is evaluated on the sequences "09" and "10" of the KITTI odometry dataset [3] following the protocol in [14]. The results are summarized in the upper half of Table 1. The protocol in Monodepth2 [4] is used to compute each of the 4 frame-to-frame transformations for evaluating on the 5-frame snippets. The results are summarized in the bottom half of Table 1. The odometry results of RM-Depth are comparable to the prior works.

Table 1. Camera pose estimation results on the KITTI odometry dataset. †Evaluation on the 5-frame trajectory is made from a combination of 4 frame-to-frame predictions. ‡The results are provided by [4]. The best result in each category is in bold.

Method	Absolute Trajectory Error (ATE)	
	Sequence "09"	Sequence "10"
Zhou *et al.* [14]	0.021 ± 0.017	0.020 ± 0.015
DF-Net [15]	0.017 ± 0.007	0.015 ± 0.009
Mahjourian *et al.* [8]	0.013 ± 0.010	0.012 ± 0.011
EPC++ [9]	0.013 ± 0.007	0.012 ± 0.008
GeoNet [13]	0.012 ± 0.007	0.012 ± 0.009
CC [10]	0.012 ± 0.007	0.012 ± 0.008
RM-Depth [6]	$\mathbf{0.0101 \pm 0.0063}$	$\mathbf{0.0096 \pm 0.0065}$
Zhou *et al.* [14]†,‡	0.050 ± 0.034	0.039 ± 0.028
Monodepth2 [4]†	$0.017 \pm \mathbf{0.008}$	$\mathbf{0.015} \pm 0.010$
RM-Depth [6] †	$\mathbf{0.0166} \pm 0.0095$	$0.0153 \pm \mathbf{0.0090}$

2. Generalization Capability

The generalization capability of RM-Depth [6] is evaluated by applying the trained model on another new dataset, Make3D [11], without fine-tuning on it. Following the protocol in Monodepth2 [4], evaluation is performed on the 134 test images in Make3D. The results are summarized in Table 2. RM-Depth performs better than the prior works. Visual results are provided in Fig. 1. Both the quantitative and qualitative results suggest that RM-Depth generalizes well to other scenes in addition to the KITTI cand Cityscapes datasets.

*This research work is not for commercial use unless a prior arrangement has been made with the author.

Table 2. Depth prediction results on the Make3D dataset. The best result in each category is in bold.

Method	AbsRel	SqRel	RMS	RMSlog
DDVO [1]	0.387	4.720	8.090	0.204
Zhou *et al.* [14]	0.383	5.321	10.470	0.478
DF-Net [15]	0.331	2.698	6.890	0.416
CC [10]	0.321	3.277	7.258	0.170
Monodepth2 [4]	0.322	3.589	7.417	0.163
RM-Depth [6]	**0.283**	**2.557**	**6.634**	**0.151**

3. Additional Results on KITTI

In Table 3, evaluation on the improved KITTI ground truth [12] as [4] is provided. RM-Depth [6] performs well comparing to the prior works.

4. Effect of Image Resolution

RM-Depth [6] is further evaluated at a higher image resolution, 1024×320. As summarized in Table 4, a high image resolution helps RM-Depth to improve the performance.

5. Complete Results on the Ablation Study

In Tables 5, 6, and 7, the full set of metrics for the experiment presented in Tables 4 to 6 of the main paper are provided.

References

[1] C.Wang, J. M. Buenaposada, R. Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, pages 2022–2030, 2018. 1, 2

[2] D. Eigen and R. Fergu. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, pages 2650–2658, 2015. 3

[3] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? In *CVPR*, pages 3354–3361, 2012. 1

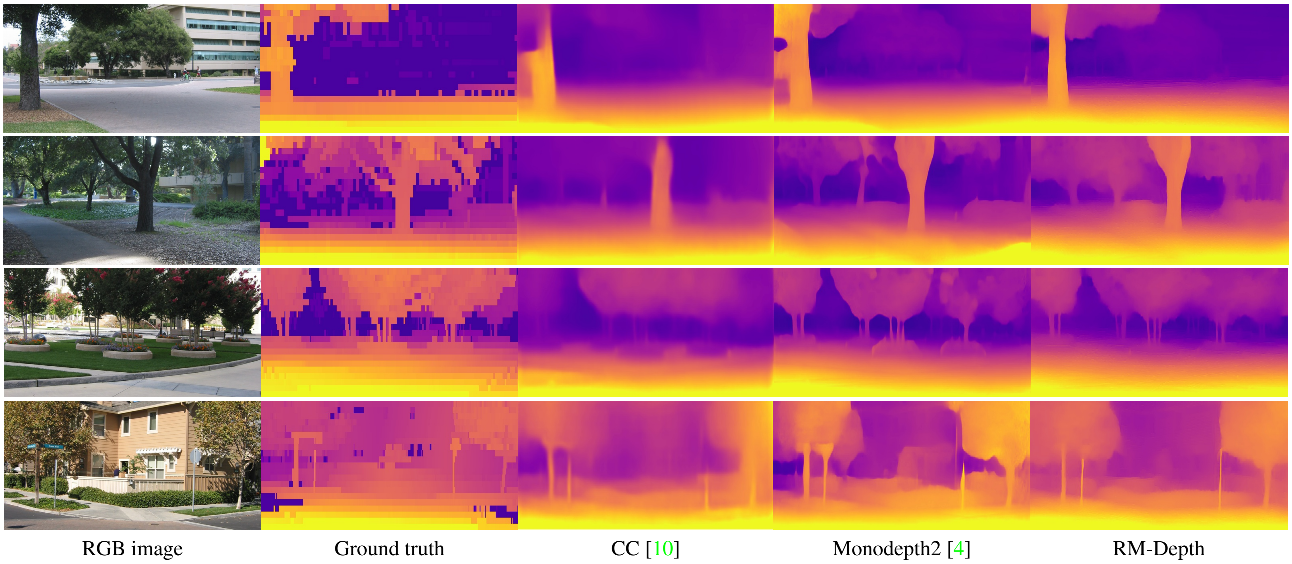

Figure 1. Examples of depth predictions on the Make3D dataset [11].

Table 3. Monocular depth results on the KITTI dataset using improved ground truth. The best result in each category is in bold.

Method	Error (lower is better)				Accuracy (higher is better)		
	AbsRel	SqRel	RMS	RMSlog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Mahjourian *et al.* [9]	0.134	0.983	5.501	0.203	0.827	0.944	0.981
GeoNet [13]	0.132	0.994	5.240	0.193	0.833	0.953	0.985
DDVO [1]	0.126	0.866	4.932	0.185	0.851	0.958	0.986
CC [10]	0.123	0.881	4.834	0.181	0.860	0.959	0.985
EPC++ [8]	0.120	0.789	4.755	0.177	0.856	0.961	0.987
Monodepth2 [4]	0.090	0.545	3.942	0.137	0.914	0.983	0.995
PackNet [5]	**0.078**	0.420	3.485	**0.121**	**0.931**	0.986	0.996
RM-Depth	0.0797	**0.373**	**3.461**	**0.121**	0.930	**0.988**	**0.997**

[4] C. Godard, O. M. Aodha, M. Firman, and G. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3828–3838, 2019. 1, 2

[5] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, pages 2485–2494, 2020. 2

[6] T.-W. Hui. RM-Depth: Unsupervised Learning of Recurrent Monocular Depth in Dynamic Scenes. In *CVPR*, pages 1675–1684, 2022. 1

[7] H. Li, A. Gordon, H. Zhao, V. Casser, and A. Angelova. Unsupervised monocular depth learning in dynamic scenes. In *CoRL*, pages 1908–1917, 2020. 3

[8] C. Luo, Z. Yang, P. Wang, Y. Wang, W. Xu, R. Nevatia, and A. Yuille. Every pixel counts ++: Joint learning of geometry and motion with 3d holistic understanding. *TPAMI*, 42(10):2624–2641, 2020. 1, 2

[9] R. Mahjourian, M. Wicke, and A. Angelovn. Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints. In *CVPR*, pages 5667–5675, 2018. 1, 2

[10] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black. Competitive Collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *CVPR*, pages 12240–12249, 2019. 1, 2

[11] A. Saxena, M. Sun, and A. Y. Ng. Make3D: Learning 3D scene structure from a single still image. *TPAMI*, 31:824–840, 2009. 1, 2

[12] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger. Sparsity invariant CNNs. In *3DV*, pages 11–20, 2017. 1

[13] Z. Yin and J. Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, pages 1983–1992, 2018. 1, 2

[14] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, pages 1851–1858, 2017. 1

[15] Y. Zou, Z. Luo, and J.-B. Huang. DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency. In *ECCV*, pages 38–55, 2018. 1

Table 4. Monocular depth results of RM-Depth using different image resolutions on the KITTI dataset by the testing split of Eigen*et al.* [2] and the Cityscapes dataset. The best result in each category is in bold.

Resolution	Testing Set	Error (lower is better)				Accuracy (higher is better)		
		AbsRel	SqRel	RMS	RMSlog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
640×192	K	0.1072	0.6870	4.4758	0.1811	0.8833	0.9637	0.9839
1024×320	K	**0.1062**	**0.6667**	**4.3002**	**0.1777**	**0.8861**	**0.9649**	**0.9847**
640×192	CS	0.0903	0.8248	5.5027	0.1430	0.9133	0.9797	0.9934
1024×320	CS	**0.0876**	**0.7549**	**5.1223**	**0.1359**	**0.9222**	**0.9827**	**0.9942**

Table 5. Ablation study of RM-Depth on KITTI. The best result in each category is in bold.

Method	Error (lower is better)				Accuracy (higher is better)		
	AbsRel	SqRel	RMS	RMSlog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
full	**0.1081**	**0.7100**	**4.5138**	**0.1831**	**0.8841**	**0.9637**	**0.9832**
w/o residual upsampling	0.1097	0.7313	4.5269	0.1839	0.8819	0.9629	0.9830
w/o RMU	0.1167	0.8186	4.7100	0.1895	0.8722	0.9604	0.9825
w/o modulation	0.1165	0.7546	4.6623	0.1910	0.8677	0.9590	0.9823
baseline (w/o my contributions)	0.1187	0.8382	4.7894	0.1927	0.8664	0.9591	0.9822

Table 6. Ablation study of RM-Depth on Cityscapes. The best result in each category is in bold.

Method	Error (lower is better)				Accuracy (higher is better)		
	AbsRel	SqRel	RMS	RMSlog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
full	**0.0903**	**0.8248**	**5.5027**	**0.1430**	0.9133	**0.9797**	**0.9934**
w/o warping	0.0933	0.9248	5.6283	0.1461	**0.9137**	0.9790	0.9925
w/o outlier-aware regularization	0.0995	0.9986	5.8281	0.1545	0.9015	0.9751	0.9916
using sparsity loss as [7]	0.1066	1.1073	6.0965	0.1642	0.8877	0.9703	0.9900
w/o object motion estimation	0.1174	1.1195	6.4542	0.1729	0.8650	0.9679	0.9906
baseline (w/o my contributions)	0.1335	1.8784	6.9748	0.1912	0.8479	0.9600	0.9856

Table 7. Ablation study of the number of RMUs in RM-Depth on KITTI. The best result in each category is in bold.

Method	Error (lower is better)				Accuracy (higher is better)		
	AbsRel	SqRel	RMS	RMSlog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
3 (L4: 1, L3: 1, L2: 1)	0.1161	0.7713	4.6799	0.1906	0.8686	0.9601	0.9825
6 (L4: 2, L3: 2, L2: 2)	0.1135	0.7490	4.6128	0.1877	0.8755	0.9612	0.9828
8 (L4: 4, L3: 2, L2: 2)	0.1098	0.7251	4.5535	0.1845	0.8809	0.9620	0.9829
13 (L4: 9, L3: 2, L2: 2)	**0.1081**	**0.7100**	**4.5138**	**0.1831**	**0.8841**	**0.9637**	**0.9832**