Supplementary Material: Pseudo RGB-D for Self-Improving Monocular SLAM and Depth Prediction

Lokender Tiwari[1[0000−0002−9217−4248]], Pan Ji[2[0000−0001−6213−554X]], Quoc-Huy Tran[2[0000−0003−1396−6544]], Bingbing Zhuang[2[0000−0002−2317−3882]], Saket Anand[1[0000−0002−6229−3940]], and Manmohan Chandraker[2,3[0000−0003−4683−2454]]

[1]IIIT-Delhi [2]NEC Labs America [3]UCSD

This supplementary material is organized as follows. We first present depth refinement results on KITTI Odometry sequences in Sec. 1. Next, we give a comparison of our pose refinement with state-of-the-art RGB SLAM approaches in Sec. 2. We further evaluate pose refinement on KITTI Leaderboard in Sec. 3. Additional implementation details and qualitative results of TUM RGB-D experiments are included in Sec. 4. Additional analysis of the self-improving loop with all the 7 depth evaluation metrics is presented in Sec. 5. Some additional qualitative depth evaluation results of KITTI Eigen experiments and pose evaluation results of KITTI Odometry experiments are presented in Sec. 6 and Sec. 7 respectively. Finally, we provide some demo videos[1] on KITTI Odometry and TUM RGB-D sequences in Sec. 8.

1 Depth Refinement Evaluation on KITTI Odometry

We evaluate the depth refinement step of our self-improving pipeline on KITTI Odometry sequences 09 and 10. The first block (*i.e.*MonoDepth2-M vs pRGBD-Refined) of the Tab. S1 shows the improved results after the depth refinement step. We also compare our method with a state-of-the-art depth refinement method DCNF [11]. Note: DCNF [11] uses *ground-truth* depths for pre-training the network, while our method uses only *unlabelled* monocular images, and still outperforms DCNF (see second block of the Tab. S1). The result shows that our self-improving framework with the wide-baseline losses (*i.e.*, symmetric depth transfer and depth consistency losses) improves the depth prediction.

2 Comparison with State-Of-The-Art SLAM Methods

In this section, we compare our pRGBD-Initial and pRGBD-Refined methods against state-of-the-art RGB SLAM methods, *i.e.*, Direct Sparse Odometry (DSO) [1], Direct Sparse Odometry with Loop Closure (LDSO) [3], and Direct Sparse Odometry in Dynamic Environments (DSOD) [7]. The results are shown in Tab. S2. From the results, it is evident that our pRGBD-Refined outperforms

[1] Demo videos: https://tiny.cc/pRGBD

Table S1. Qualitative depth evaluation on KITTI Odometry sequences 09 and 10. M: self-supervised monocular supervision for fine-tuning. '-' means the result is not available from the paper. Our results are after *5 self-improving loops*. Note: DCNF [11] uses *ground-truth* depths for pre-training. Best results in each block is in **bold**.

		Depth	**Lower is better**				**Higher is better**		
Method	Train	Cap	Abs Rel	Sq Rel	RMSE	RMSE $\log_2$	a1	a2	a3
MonoDepth2-M [5]	M	80	0.123	0.703	4.165	0.188	0.854	0.956	0.985
pRGBD-Refined	M	80	**0.121**	**0.649**	**3.995**	**0.184**	**0.853**	**0.960**	**0.986**
DCNF [11]	M	20	0.112	-	2.047	-	-	-	-
pRGBD-Refined	M	20	**0.098**	**0.242**	**1.610**	**0.145**	**0.906**	**0.978**	**0.993**

all the competing methods in Absolute Trajectory Error (RMSE) and Relative Translation (Rel Tr) Error . While the improvement in Absolute Trajectory Error (RMSE) and Relative Translation (Rel Tr) error is substantial, the performance in Relative Rotation (Rel Rot) is not comparable. The higher Rel Rot errors of our method compared to other RGB ORB-SLAM methods could be due to the high uncertainty of CNN-predicted depths for far-away points, which affects our rotation estimation [6]. However, if we compare Rel Rot error of pRGBD-Initial with the pRGBD-Refined, as depth prediction improves (see Tab. S1 MonoDepth2-M/pRGBD-Initial vs pRGBD-Refined) the Rel Rot error also improves (see Tab. S2).

Table S2. Comparison with state-of-the-art RGB SLAM methods on KITTI Odometry sequences 09 and 10. Here, - means the result is not available from the original paper. $*$ denotes the result is obtained from [7].

	Seq. 09			Seq. 10		
Method	RMSE	Rel Tr	Rel Rot	RMSE	Rel Tr	Rel Rot
RGB ORB-SLAM[8]	18.34	7.42	0.004	8.90	5.85	0.004
DSO[1]	74.29	72.27*	**0.002***	16.32	80.81*	**0.002***
LDSO[3]	21.64	-	-	17.36	-	-
DSOD[7]	-	13.85	**0.002**	-	13.53	**0.002**
pRGBD-Initial	12.21	4.26	0.011	8.30	5.55	0.017
pRGBD-Refined	**11.97**	**4.20**	0.010	**6.35**	**4.40**	0.016

3 KITTI Odometry Leaderboard Results

In the main paper, we keep the default setting from ORB-SLAM, which leads to tracking failures of all methods in a few sequences (i.e., see Tab. 3 of the main paper). The KITTI Odometry leaderboard requires the results of all sequences (i.e., sequences 11-21) for evaluation. Therefore, we increase the minimum number of inliers for adding keyframes from 100 to 500 so that our pRGBD-Refined succeeds on all sequences. We report the results of our pRGBD-Refined on the KITTI Odometry leaderboard in Tab. S3. Results show our method outperforms the

competing monocular/LiDAR-based methods both in terms of relative translation and rotation errors.

Table S3. Quantitative pose evaluation results on KITTI Odometry leaderboard. Note that we use the estimated trajectories from ORB-SLAM2-S [8] for global scale alignment. The best performance is in **bold**.

Method	Rel Tr	Rel Rot
ORB-SLAM2-S [8]	1.70	0.0028
OABA [2]	20.95	0.0135
VISO2-M [4]	11.94	0.0234
BLO [10]	9.21	0.0163
VISO2-M+GP [4, 9]	7.46	0.0245
pRGBD-Refined	**6.24**	**0.0097**

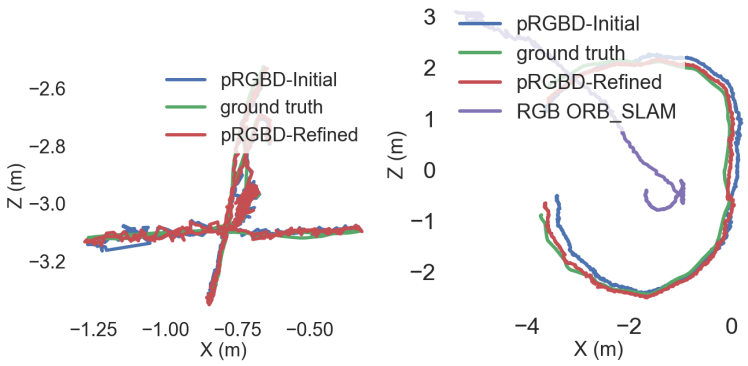
4 Experiments on TUM RGB-D Sequences

4.1 Implementation Details

We pre-train/fine-tune the depth network on image resolution 480×320. For pre-training, we set the learning rate to 10^{-4} initially, reduce it to 10^{-5} after 20 epochs, and train for 30 epochs. For fine-tuning, we extract camera poses, 2D keypoints and the associated depths from keyframes while running RGB-D ORB-SLAM on the training sequences. We fine-tune the depth network with the fixed learning rate of 10^{-6}. We use the following 6 sequences for pre-training/fine-tuning: 1. fr3/long_office_household, 2. fr3/long_office_household_validation, 3. fr3/sitting_xyz, 4. fr3/structure_texture_far, 5. fr3/structure_texture_near, 6. fr3/teddy, and the following 2 sequences for testing: 1. fr3/walking_xyz, 2. fr3/large_cabinet_validation. Note that these are the only 8 sequences with provided rectified images among the entire TUM RGB-D dataset.

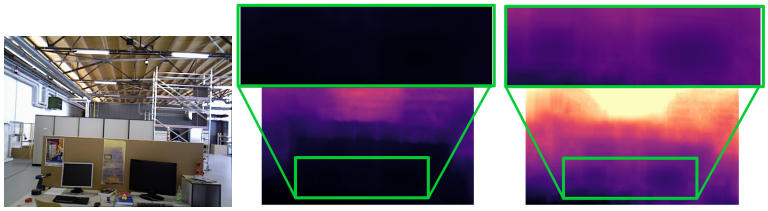
4.2 Qualitative Results

Fig. S1(a) and Fig. S1(b) shows qualitative pose evaluation results on test sequences *walking_xyz* and *large_cabinet_validation* respectively. The results, show the increased robustness and accuracy by pRGBD-Refined. In particular, RGB ORB-SLAM fails on walking_xyz, while pRGBD-Refined succeeds and achieves the best performance on both sequences. Some qualitative depth refinement results are presented in Fig. S2. It can be seen that the disparity between the depth values of nearby and farther scene points become clearer, *e.g.*, see depth around the two monitors.



(a) fr3/walking_xyz (b)fr3/large_cabinet_validation

Fig. S1. Qualitative pose evaluation results on TUM RGB-D sequences. Note that RGB ORB-SLAM fails in (a).



(a) RGB (b) pRGBD-Initial (c) pRGBD-Refined

Fig. S2. Qualitative depth evaluation results on TUM RGB-D sequences.

5 Additional Plots of Self-Improving Loop Analysis

In the main paper, we have shown behaviours of 3 depth evaluation metrics named as (Sq. Rel), (RMSE) and (a2). In this section we present behaviours of all 7 metrics and pose evaluation metrics. Our analysis in the Sec. 5 of the main paper holds true with respect to all the 7 depth evaluation metrics.

6 Additional Depth Refinement Qualitative Results

Fig. S4 shows some visual improvements in depth predictions of farther scene points. Fig. S5 shows some additional qualitative results, where pRGBD-Refined shows visible improvements at occlusion boundaries and thin objects. The reason for the improvements is the aggregated cues from multiple views with wider baselines (e.g., our depth transfer and depth consistency losses) lead to more well-posed depth recovery.

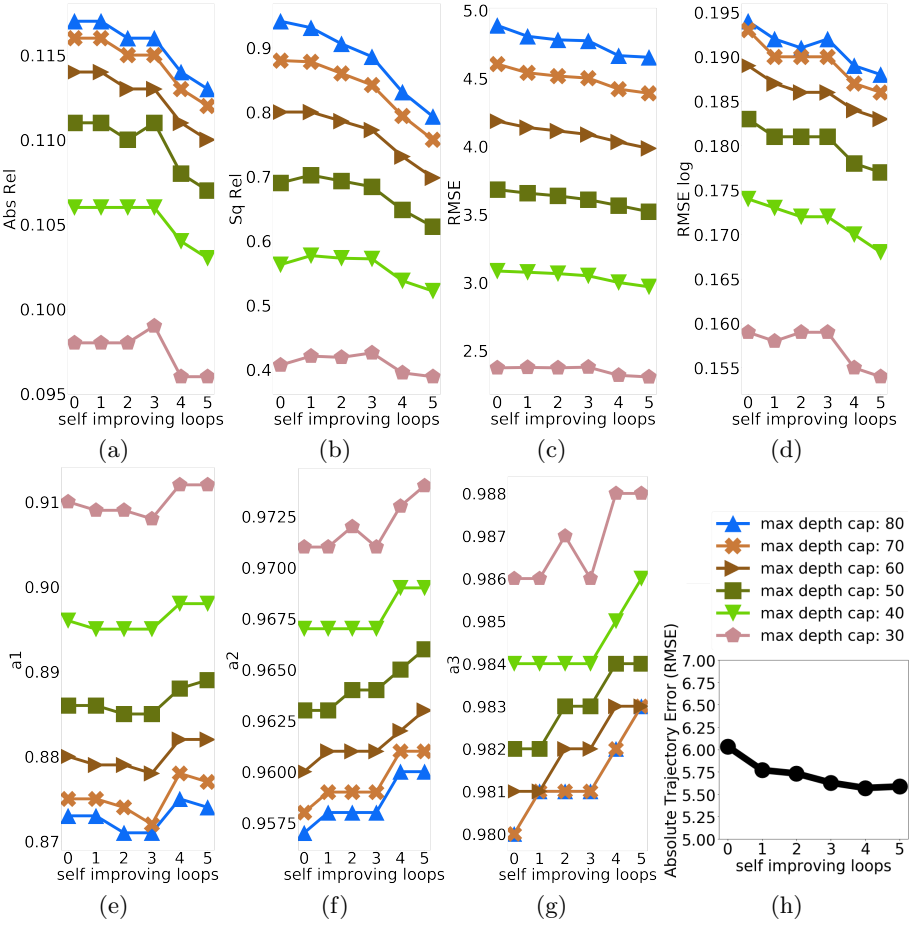


Fig. S3. Depth/Pose evaluation metrics w.r.t. self-improving loops. (a). Absolute Relative (Abs Rel) (*lower is better*) (b). Squared Relative (Sq Rel) (*lower is better*) (c). RMSE (*lower is better*) (d). RMSE Log (*lower is better*), (e). a1 (*higher is better*), (f). a2 (*higher is better*), (g). a3 (*higher is better*) and (h) Absolute Trajectory Error (RMSE) (*lower is better*). Depth evaluation metrics in (a-g) are computed at different max depth caps ranging from 30-80 meters.

7 Additional Pose Refinement Qualitative Results

Some additional pose refinement qualitative results are shown in Fig. S6. In all the three sequences our pRGBD-Refined aligned well with the ground-truth trajectory. Note that both RGB ORB-SLAM and our pRGBD-Initial fail on sequence 12, whereas our pRGBD-Refined succeeds, showing the enhanced robustness by our self-improving framework.

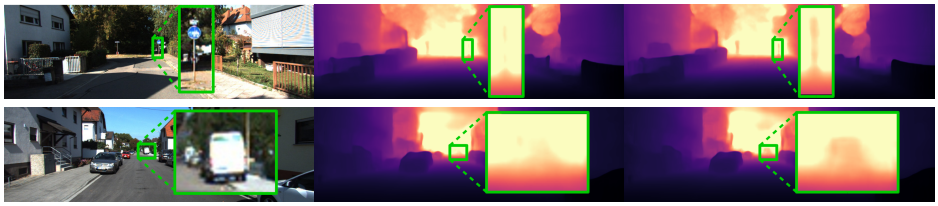


Fig. S4. Qualitative depth evaluation results on KITTI Odometry test set. Improvement in depth prediction of farther scene points.

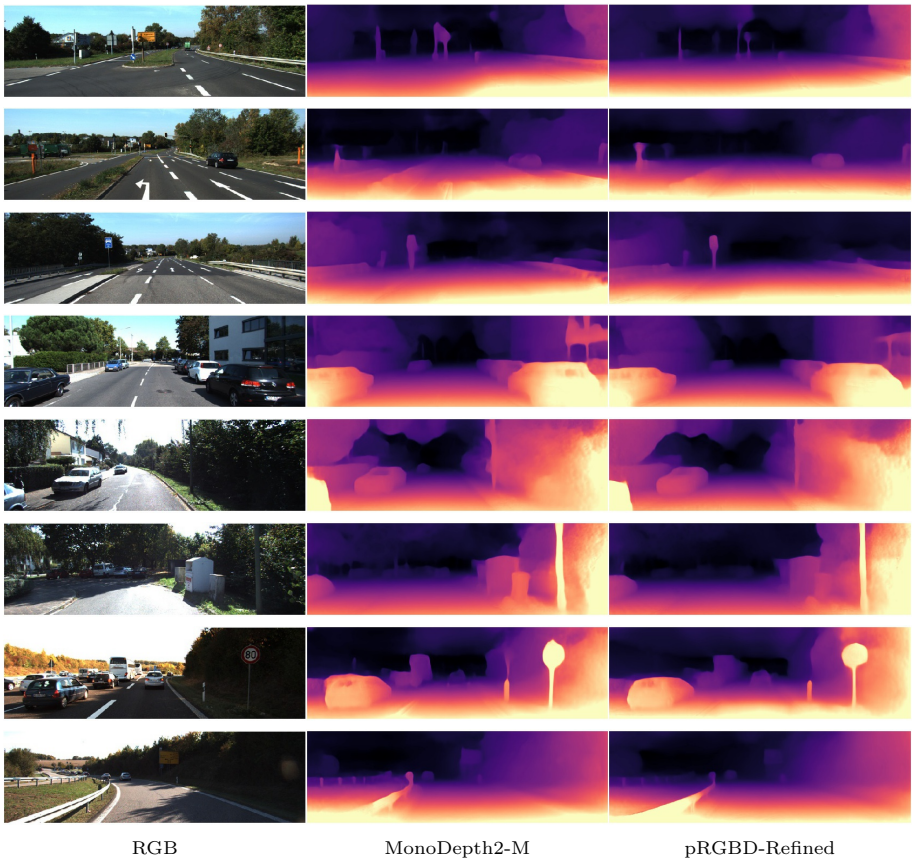
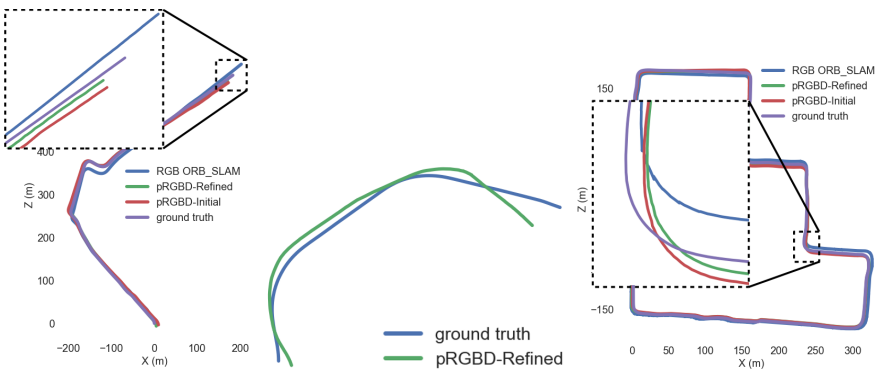


Fig. S5. Qualitative depth evaluation results on KITTI Raw Eigen split test set. MonoDepth2-M: MonoDepth2 trained using monocular images,

8 Demo Videos

We include example videos [2] on sequences 11 and 19 of KITTI Odometry (i.e., *kitti_seq_11.mp4* and *kitti_seq_19.mp4*, respectively) and sequence fr3/large_cabinet_

[2] Demo videos: https://tiny.cc/pRGBD



(a) Seq 11 (b) Seq 12 (a) Seq 15

Fig. S6. Qualitative pose evaluation results on KITTI Odometry sequences. Note that both RGB ORB-SLAM and pRGBD-Initial fail in (b).

validation of TUM RGB-D (i.e., *tum_large_cabinet_validation.mp4*). In particular, we illustrate the improvements in depth prediction at frames 140, 352 of *kitti_seq_11.mp4*, frames 1652, 3248, 3529 of *kitti_seq_19.mp4*, and frames 153, 678 of *tum_large_cabinet_validation.mp4*. In addition, we highlight the failure of RGB ORB-SLAM at frame 2985 of *kitti_seq_19.mp4*.

References

1. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(3), 611–625 (2017)
2. Frost, D.P., Kähler, O., Murray, D.W.: Object-aware bundle adjustment for correcting monocular scale drift. In: ICRA (2016)
3. Gao, X., Wang, R., Demmel, N., Cremers, D.: Ldso: Direct sparse odometry with loop closure. In: IROS (2018)
4. Geiger, A., Ziegler, J., Stiller, C.: Stereoscan: Dense 3d reconstruction in real-time. In: IV (2011)
5. Godard, C., Aodha, O.M., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: ICCV (2019)
6. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge University Press (2003)
7. Ma, P., Bai, Y., Zhu, J., Wang, C., Peng, C.: Dsod: Dso in dynamic environments. IEEE Access **7**, 178300–178309 (2019)
8. Mur-Artal, R., Tardós, J.D.: ORB-SLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Transactions on Robotics **33**(5), 1255–1262 (2017)
9. Song, S., Chandraker, M.: Robust scale estimation in real-time monocular sfm for autonomous driving. In: CVPR (2014)
10. Velas, M., Spanel, M., Hradis, M., Herout, A.: Cnn for imu assisted odometry estimation using velodyne lidar. In: IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC) (2018)

11. Yin, X., Wang, X., Du, X., Chen, Q.: Scale recovery for monocular visual odometry using depth estimated with deep convolutional neural fields. In: ICCV (2017)