GeoRefine: Self-Supervised Online Depth Refinement for Accurate Dense Mapping – Supplementary Material

Pan Ji⋆, Qingan Yan⋆, Yuxin Ma, and Yi Xu

OPPO US Research Center, InnoPeak Technology, Inc

1 Implementation Details

In this section, we describe the implementation details of RAFT-SLAM and present a simple mechanism to handle SLAM failures.

1.1 RAFT-SLAM

Our system utilizes ROS as the agent for cross-language communication. Consecutive frames are fed into the RAFT network [22] to get pair-wise flow predictions, including both the forward and the backward flows. For all our experiments, we use the RAFT flow model that is pretrained on FlyingThings3D, *i.e.*, raft-things.pth downloaded from https://github.com/princeton-vl/RAFT. In the monocular mode, after the system successfully initializes, we continuously align the map points and camera poses to CNN depth for five steps to make their scales consistent to each other.

1.2 Model Selection

We choose DPT as the main baseline for depth refinement due to two reasons: 1) it's one of the most recent works (in ICCV'21); 2) it has exceptional generalizability so that our GeoRefine can be deployed in any unseen environments without additional finetuning. It is also feasible to adopt other benchmark algorithms. For instance, we experiment with a more recent baseline DNet [1] and report results on two randomly selected sequences from the ScanNet test set in Tab. 7. We can see that our GeoRefine achieves consistent improvement over this baseline as well.

1.3 SLAM Failures

It is hard to ensure RAFT-SLAM never encounters failure cases. We observe that it fails occasionally on sequences with strong motion blur and significant rolling-shutter artifacts. In the event of SLAM failures, we want the depth model to be rarely disrupted and the system is supposed to continue to run after the

⋆ Joint first authorship. P. Ji is the corresponding author (peterji530@gmail.com).

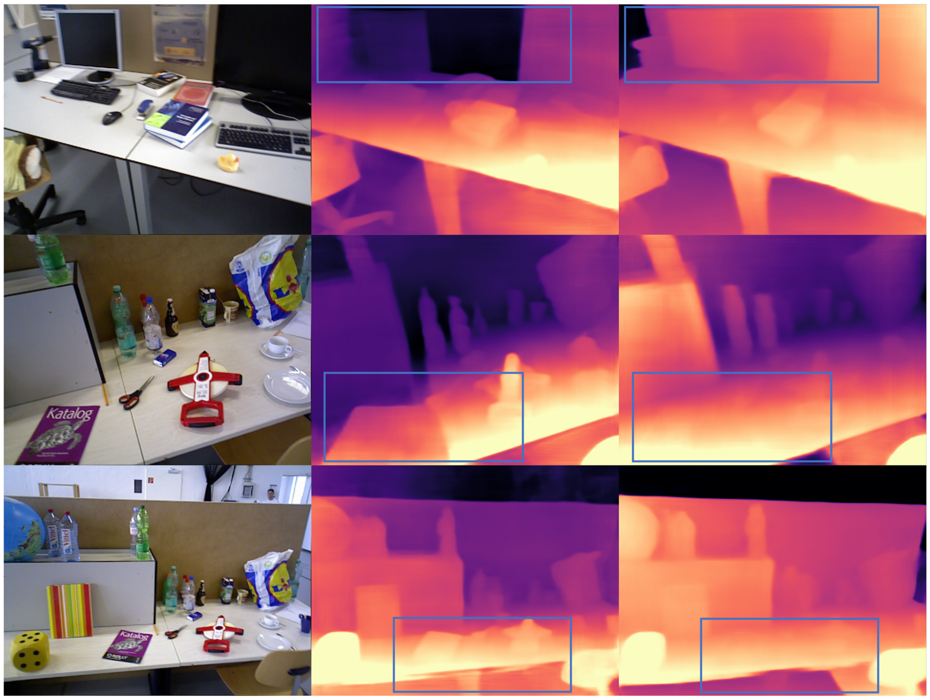


Fig. 1: Qualitative results on TUM-RGBD. From left to right: input images, depth maps by DPT, depth maps by our GeoRefine. Our method is able to eliminate many artifacts and erroneous predictions compared to DPT.

SLAM module recovers. To this end, we employ a simple strategy, *i.e.*, after the depth refinement module receives a signal of SLAM failure, the system clears the queues both for keyframe and per-frame data. In this case, the keyframe depth refinement process is paused, but the per-frame depth inference can still run if depth maps for all frames are demanded.

2 EuRoC

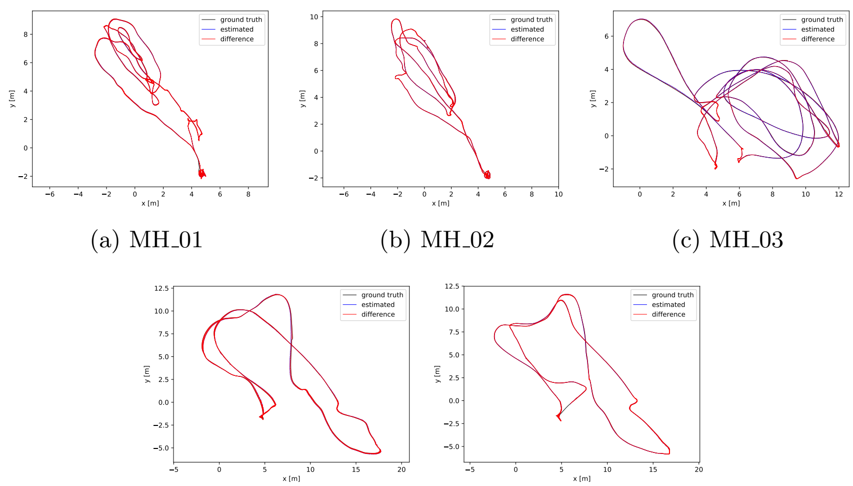
In this section, we include additional depth and pose results on EuRoC. More qualitative results can be found in the attached videos.

2.1 GeoRefine-MD2

We present depth results of GeoRefine using a self-supervised model, *i.e.*, Monodepth2 [8], as the base model on EuRoC. We take monocular and stereo images from five sequences (MH_01, MH_02, MH_04, V1_01, and V1_02) as the training set to train the base model Monodepth2. Since stereo images with a known baseline distance are used, the pretrained Monodepth2 is scale-aware. The quantitative depth results are shown in Tab. 3, from which we can see that our system,

Table 1: pRGBD SLAM results on EuRoC (RMSE ATE in meters).

Method	MH_01	MH_02	MH_03	MH_04	MH_05	V1_01	V1_02	V1_03	V2_01	V2_02	V2_03	Mean
ORB-SLAM3 [2]	**0.016**	0.027	**0.028**	0.138	0.072	**0.033**	**0.015**	0.033	0.023	0.029	x	-
Ours-pRGBD	0.025	**0.023**	0.031	**0.064**	**0.060**	**0.033**	**0.015**	**0.023**	**0.022**	**0.016**	**0.034**	**0.031**



(a) MH_01 (b) MH_02 (c) MH_03

(d) MH_04 (e) MH_05

Fig. 2: Visualized trajectory results on EuRoC MH sequences. Best viewed on screen with zoom-in.

Table 2: Ablation study on EuRoC Sequence V2_03 in pRGBD mode

Method	Depth						Odometry
	MAE ↓	Abs Rel ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	RMSE ATE ↓
DPT [20]	0.283	0.099	0.366	0.905	0.979	0.994	-
Our BaseSystem	0.216	0.076	0.288	0.933	0.989	0.998	0.176
+ Refined Depth	0.199	0.065	0.268	0.958	0.995	**0.999**	0.133
+ RAFT-flow	0.171	0.056	0.237	0.972	0.995	0.998	0.069
+ Remove BA Term	**0.152**	**0.051**	**0.214**	**0.975**	**0.997**	**0.999**	**0.034**

denoted as “Ours-MD2”, improves over Monodepth2 by a significant margin in all three SLAM modes.

2.2 Odometry and Ablation

Tab. 1 and Tab. 2 report the odometry results of our proposed RAFT-SLAM in the pRGBD mode and the corresponding ablation study. It’s evident that our pRGBD RAFT-SLAM outperforms the baseline, *i.e.*, ORB-SLAM3, both in terms of robustness and accuracy, and each proposed new component contributes to the improvement. Note that “Our BaseSystem” uses only the pretrained depth from DPT to form a pRGBD mode. Fig. 2 shows the visualized trajectories on EuRoC MH sequences.

Table 3: Quantitative depth evaluation on EuRoC using Monodepth2.

Method	Monocular				Visual-Inertial				pRGBD			
	MAE ↓	AbsRel ↓	RMSE ↓	δ_1 ↑	MAE ↓	AbsRel ↓	RMSE ↓	δ_1 ↑	MAE ↓	AbsRel ↓	RMSE ↓	δ_1 ↑
V1_03												
Monodepth2 [8]	0.305	0.111	0.413	0.886	0.360	0.132	0.464	0.815	0.305	0.111	0.413	0.886
Ours-MD2	**0.184**	**0.066**	**0.272**	**0.960**	**0.178**	**0.062**	**0.255**	**0.972**	**0.178**	**0.059**	**0.251**	**0.966**
V2_01												
Monodepth2 [8]	0.423	0.153	0.581	0.800	0.490	0.181	0.648	0.730	0.423	0.153	0.581	0.800
Ours-MD2	**0.202**	**0.063**	**0.306**	**0.960**	**0.169**	**0.059**	**0.265**	**0.968**	**0.191**	**0.060**	**0.295**	**0.958**
V2_02												
Monodepth2 [8]	0.597	0.191	0.803	0.723	0.769	0.233	0.963	0.562	0.597	0.191	0.803	0.723
Ours-MD2	**0.218**	**0.065**	**0.350**	**0.955**	**0.193**	**0.060**	**0.320**	**0.964**	**0.199**	**0.059**	**0.327**	**0.962**
V2_03												
Monodepth2 [8]	0.601	0.211	0.784	0.673	0.764	0.258	0.912	0.498	0.601	0.211	0.784	0.673
Ours-MD2	**0.192**	**0.064**	**0.266**	**0.956**	**0.171**	**0.059**	**0.251**	**0.968**	**0.207**	**0.069**	**0.297**	**0.951**

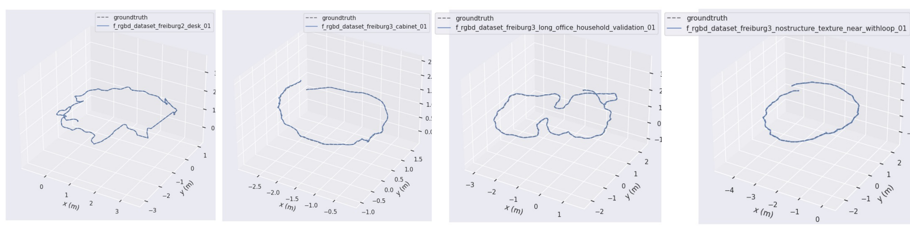


Fig. 3: Qualitative pose results of our system under the pRGBD mode on TUM-RGBD. Best viewed on screen with zoom-in.

Table 4: Odometry results on TUM-RGBD in terms of **RPE [m/s]**. "X" means no pose output due to system failure and "(X)" means partial pose results.

Method	f2/desk	f2/pio_360	f2/pio_slam	f3/cbnet	f3/l_o_h_val	f3/ns_t_nr_lp	f3/str_nt_f	f3/str_nt_n	mean
ORB-SLAM3 [2]	0.039	0.155(X)	X	0.160(X)	0.024	0.604	X	X	-
Li [13]	0.158	0.201	0.176	0.213	0.133	0.159	0.104	0.207	0.169
Ours-Mono	**0.025**	**0.075**	0.161	0.079	**0.022**	**0.028**	0.107	0.195(X)	0.089
Ours-pGRBD	0.033	0.092	**0.133**	**0.023**	0.028	0.031	**0.042**	**0.092**	**0.059**

3 TUM-RGBD

We evaluate our GeoRefine on a few more sequences from the TUM-RGBD dataset. We adopt the same settings as in the main paper and use the DPT model [20] pretrained on NYUv2 as our initial model. The quantitative depth results are shown in Tab. 5, from which we can observe consistent and significant improvements by our GeoRefine over the pretrained model. Qualitative results can be found in Fig. 1, Fig. 5 and the attached video.

In addition, we compare with [13] and show odometry results in terms of relative pose error (RPE) on TUM-RGBD in Tab. 4. Compared to the baseline ORB-SLAM3 [2], the improved odometry results by our system verify that *using RAFT makes the SLAM system more robust and accurate*. In particular, our method in both the monocular and pRGBD modes outperforms a recent deep odometry method [13] by a significant margin. See Fig. 3 for qualitative pose results of our system under the pRGBD mode.

Table 5: Quantitative depth evaluation on additional TUM-RGBD sequences.

Method	Monocular						pRGBD					
	MAE ↓	AbsRel ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	MAE ↓	AbsRel ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
freiburg3_long_office_household												
DPT [20]	0.366	0.129	0.762	0.833	0.926	0.955	0.366	0.129	0.762	0.833	0.926	0.955
Ours-DPT	**0.175**	**0.078**	**0.349**	**0.926**	**0.973**	**0.993**	**0.146**	**0.065**	**0.315**	**0.947**	**0.989**	**0.997**
freiburg3_long_office_household_validation												
DPT [20]	0.350	0.136	0.750	0.836	0.924	0.948	0.350	0.136	0.750	0.836	0.924	0.948
Ours-DPT	**0.171**	**0.078**	**0.380**	**0.930**	**0.965**	**0.976**	**0.151**	**0.071**	**0.341**	**0.941**	**0.977**	**0.993**
freiburg3_nostructure_texture_near_withloop												
DPT [20]	0.129	0.103	0.163	0.914	0.999	1.000	0.129	0.103	0.163	0.914	0.999	1.000
Ours-DPT	**0.028**	**0.024**	**0.039**	**0.996**	**1.000**	**1.000**	**0.028**	**0.024**	**0.039**	**1.000**	**1.000**	**1.000**

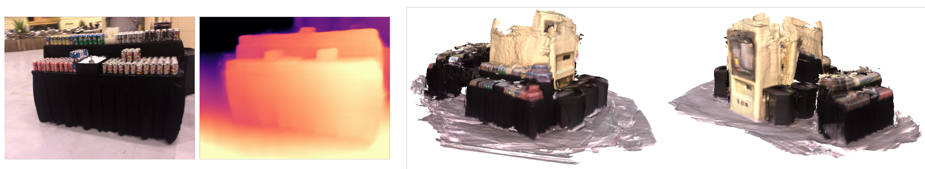


Fig. 4: Global reconstruction on ScanNet (scene0228_00) using the refined depth maps by GeoRefine.

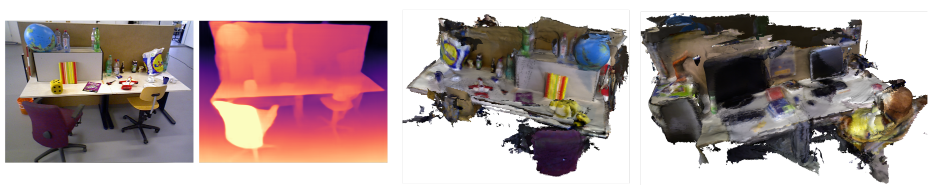


Fig. 5: Global reconstruction on TUM-RGBD (freiburg3_long_office_household) using the refined depth maps by GeoRefine.

4 ScanNet

ScanNet [4] is an indoor RGB-D dataset consisting of more than 1500 scans. This dataset was captured by a handheld device, so motion blur exists in most of the sequences, posing challenges both for monocular SLAM and depth refinement. Moreover, camera translations in this dataset are small as most of the sequences are from small rooms (*e.g.*, bathrooms and bedrooms). To test our GeoRefine, we sample three sequences that have relatively larger camera translations and run our system using NYUv2-pretrained DPT [20] as the base model. The results are summarized in Tab. 6. The pretrained DPT model performs well on ScanNet, reaching *Abs Rel* of 6.3% to 8.0%, probably due to dataset similarity between ScanNet and NYUv2. Our GeoRefine continues to improve the depth results in most of the metrics. In particular, on scene0228_00, our system reduces *Abs Rel* from 8.0% to 5.0% and increases δ_1 from 93.1% to 97.9%. Qualitative results can be found in Fig. 4 and the attached video.

Without loss of generality, we also experiment with a different baseline DNet [1] for online depth refinement and conduct comparisons with both its monocu-

Table 6: Quantitative depth evaluation on ScanNet.

Method	Monocular						pRGBD					
	MAE ↓	AbsRel ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	MAE ↓	AbsRel ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
scene0084_00												
DPT [20]	0.118	0.072	0.164	0.959	0.994	**0.999**	0.118	0.072	0.164	0.959	**0.994**	0.999
Ours-DPT	**0.099**	**0.062**	**0.137**	**0.967**	0.993	**1.000**	**0.089**	**0.052**	**0.145**	**0.983**	**0.995**	0.997
scene0228_00												
DPT [20]	0.205	0.080	0.380	0.931	0.986	0.998	0.205	0.080	0.380	0.931	0.986	0.998
Ours-DPT	**0.132**	**0.050**	**0.272**	**0.979**	**0.996**	**0.999**	**0.141**	**0.051**	**0.361**	**0.980**	**0.996**	**0.998**
scene0451_05												
DPT [20]	0.184	0.080	0.252	0.947	**0.997**	**1.000**	0.184	0.080	0.252	0.947	**0.997**	**1.000**
Ours-DPT	**0.164**	**0.065**	**0.248**	**0.961**	0.995	0.999	**0.153**	**0.061**	**0.237**	**0.967**	0.996	0.999

lar and multi-view stereo (MVS) models (MaGNet) on two randomly selected ScanNet test sequences, which is reported in Tab. 7. Note that, different from MVS methods, our system uses multi-views only in the losses, not the input to depth models. Therefore, GeoRefine is still a monocular-based method. Comparing with MVS methods is only to illustrate its robustness. As known that MVS methods highly rely on perfect poses which are not always available in practice, to verify this, we test two versions of MaGNet: one with groundtruth poses (denoted as "+GtPose") and the other with poses from our GeoRefine (as "+OurPose"). We can see that MaGNet suffers from a notable performance drop under "+OurPose" and our strategy outperforms both versions of MaGNet.

Table 7: Depth evaluation on the ScanNet test set.

Scene	Method	MAE ↓	AbsRel ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
0782_00	DNet (Mono) [1]	0.283	0.088	0.392	0.914	0.993	**1.000**
	MaGNet (MVS) [1]+GtPose	0.223	0.072	0.323	0.946	0.995	**1.000**
	MaGNet (MVS) [1]+OurPose	0.299	0.098	0.387	0.915	0.991	**1.000**
	Ours-DNet-pRGBD	**0.132**	**0.046**	**0.202**	**0.981**	**0.997**	**1.000**
0793_00	DNet (Mono) [1]	0.232	0.083	0.331	0.933	0.993	**0.999**
	MaGNet (MVS) [1]+GtPose	0.154	0.056	0.239	0.972	**0.997**	**0.999**
	MaGNet (MVS) [1]+OurPose	0.229	0.085	0.324	0.928	0.991	**0.999**
	Ours-DNet-pRGBD	**0.145**	**0.052**	**0.232**	**0.976**	0.996	**0.999**

5 KITTI

We show the depth results on KITTI in Tab. 8. The motion threshold for keyframes (or per-frame) is set to 0.25 m (or 0.05 m), λ_m to 0.01, and three frames (*i.e.*, 0, -1, 1) are used to build the loss; other parameters remain the same as in the main paper. Compared to the base model Monodepth2, our GeoRefine reduces *Abs Rel* by 1% and improves δ_1 by 2.8%. However, due to moving objects in KITTI, the improvement by our system is not as significant as in non-dynamic indoor environments.

Table 8: Depth evaluation results on the KITTI Eigen split test set. M: self-supervised monocular supervision; S: self-supervised stereo supervision; D: depth supervision; Align: scale alignment; Y: Yes; N: No. '-' means the result is not available from the paper. Best numbers in each block is marked in bold.

	Method	Train	Align	Error Metric				Accuracy Metric		
				Abs Rel	Sq Rel	RMSE	RMSE log	δ_1	δ_2	δ_3
Supervised	Eigen [5]	D	N	0.203	1.548	6.307	0.282	0.702	0.890	0.890
	Liu [14]	D	N	0.201	1.584	6.471	0.273	0.680	0.898	0.967
	Kuznietsov [11]	DS	N	0.113	0.741	4.621	0.189	0.862	0.960	0.986
	SVSM FT [15]	DS	N	0.094	0.626	4.252	0.177	0.891	0.965	0.984
	Guo [9]	DS	N	0.096	0.641	4.095	0.168	0.892	0.967	0.986
	DORN [6]	D	N	**0.072**	**0.307**	**2.727**	**0.120**	**0.932**	**0.984**	**0.994**
Self-Supervised	Yang [28]	M	Y	0.182	1.481	6.501	0.267	0.725	0.906	0.963
	Mahjourian [17]	M	Y	0.163	1.240	6.220	0.250	0.762	0.916	0.968
	Klodt [10]	M	Y	0.166	1.490	5.998	-	0.778	0.919	0.966
	DDVO [25]	M	Y	0.151	1.257	5.583	0.228	0.810	0.936	0.974
	GeoNet [29]	M	Y	0.149	1.060	5.567	0.226	0.796	0.935	0.975
	DF-Net [31]	M	Y	0.150	1.124	5.507	0.223	0.806	0.933	0.973
	Ranjan [21]	M	Y	0.148	1.149	5.464	0.226	0.815	0.935	0.973
	EPC++ [15]	M	Y	0.141	1.029	5.350	0.216	0.816	0.941	0.976
	Struct2depth(M) [3]	M	Y	0.141	1.026	5.291	0.215	0.816	0.945	0.979
	WBAF [30]	M	Y	0.135	0.992	5.288	0.211	0.831	0.942	0.976
	pRGBD-Refined [23]	M	Y	0.113	0.793	4.655	**0.188**	0.874	0.960	0.983
	Luo [16]	M	Y	0.130	2.086	4.876	0.205	0.878	0.946	0.970
	Li [13]	M	Y	0.106	**0.701**	**4.129**	0.210	0.889	**0.967**	**0.984**
	Garg [7]	S	N	0.152	1.226	5.849	0.246	0.784	0.921	0.967
	3Net (R50) [19]	S	N	0.129	0.996	5.281	0.223	0.831	0.939	0.974
	Monodepth2-S [8]	S	N	0.109	0.873	4.960	0.209	0.864	0.948	0.975
	SuperDepth [18]	S	N	0.112	0.875	4.958	0.207	0.852	0.947	0.977
	monoResMatch [24]	S	N	0.111	0.867	4.714	0.199	0.864	0.954	0.979
	DepthHints [26]	S	N	0.106	0.780	4.695	0.193	0.875	**0.958**	**0.980**
	DVSO [27]	S	N	**0.097**	**0.734**	**4.442**	**0.187**	**0.888**	**0.958**	**0.980**
	UnDeepVO [12]	MS	N	0.183	1.730	6.570	0.268	-	-	-
	EPC++ [15]	MS	N	0.128	0.935	5.011	0.209	0.831	0.945	0.979
	Monodepth2 [8]	MS	N	0.106	0.818	4.750	0.196	0.874	0.957	0.979
	Ours-MD2-Mono	(S)M	Y	**0.096**	0.766	4.436	**0.177**	**0.902**	**0.963**	**0.982**

6 Runtime

Our RAFT-SLAM and online dense mapping modules run in parallel with a rough 1 fps runtime in total. On the RAFT-SLAM side, since we only publish one pair image each time to the RAFT network end in a down-scaled resolution, the per-frame tracking can be executed at 5 fps. For dense mapping, the per-frame refinement step runs efficiently with around 10 fps when using the pretrained Monodepth2 model in a lower resolution and or using the pretrained DPT model. Keyframe refinement is the most time-consuming step in our system, costing around 300 ms each time. The rest of runtime is consumed by data loading, pre-processing, and cross-module communication, which can be further optimized in a future version.

References

1. et al., G.: Multi-view depth estimation by fusing single-view depth probability with multi-view geometry. CVPR (2022)
2. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: Orb-slam3: An accurate open-source library for visual, visual-inertial and multi-map slam. arXiv preprint arXiv:2007.11898 (2020)
3. Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In: AAAI (2019)
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR. pp. 5828–5839 (2017)
5. 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)
6. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: CVPR (2018)
7. Garg, R., BG, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: ECCV (2016)
8. Godard, C., Aodha, O.M., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: ICCV (2019)
9. Guo, X., Li, H., Yi, S., Ren, J., Wang, X.: Learning monocular depth by distilling cross-domain stereo networks. In: ECCV (2018)
10. Klodt, M., Vedaldi, A.: Supervising the new with the old: learning sfm from sfm. In: ECCV (2018)
11. Kuznietsov, Y., Stückler, J., Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. In: CVPR (2017)
12. Li, R., Wang, S., Long, Z., Gu, D.: Undeepvo: Monocular visual odometry through unsupervised deep learning. In: ICRA (2018)
13. Li, S., Wu, X., Cao, Y., Zha, H.: Generalizing to the open world: Deep visual odometry with online adaptation. In: CVPR. pp. 13184–13193 (2021)
14. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. TPAMI **38**(10), 2024–2039 (2015)
15. 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. arXiv preprint arXiv:1810.06125 (2018)
16. Luo, X., Huang, J.B., Szeliski, R., Matzen, K., Kopf, J.: Consistent video depth estimation. TOG **39**(4), 71–1 (2020)
17. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In: CVPR (2018)
18. Pillai, S., Ambruş, R., Gaidon, A.: Superdepth: Self-supervised, super-resolved monocular depth estimation. In: ICRA (2019)
19. Poggi, M., Tosi, F., Mattoccia, S.: Learning monocular depth estimation with unsupervised trinocular assumptions. In: 3DV (2018)
20. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV. pp. 12179–12188 (2021)
21. Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M.J.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: CVPR. pp. 12240–12249 (2019)

22. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV. pp. 402–419 (2020)
23. Tiwari, L., Ji, P., Tran, Q.H., Zhuang, B., Anand, S., Chandraker, M.: Pseudo rgb-d for self-improving monocular slam and depth prediction. In: ECCV. pp. 437–455 (2020)
24. Tosi, F., Aleotti, F., Poggi, M., Mattoccia, S.: Learning monocular depth estimation infusing traditional stereo knowledge. In: CVPR (2019)
25. Wang, C., Buenaposada, J.M., Zhu, R., Lucey, S.: Learning depth from monocular videos using direct methods. In: CVPR. pp. 2022–2030 (2018)
26. Watson, J., Firman, M., Brostow, G.J., Turmukhambetov, D.: Self-supervised monocular depth hints. In: ICCV (2019)
27. Yang, N., Wang, R., Stückler, J., Cremers, D.: Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In: ECCV (2018)
28. Yang, Z., Wang, P., Xu, W., Zhao, L., Nevatia, R.: Unsupervised learning of geometry with edge-aware depth-normal consistency. In: AAAI (2018)
29. Yin, Z., Shi, J.: GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In: CVPR (2018)
30. Zhou, L., Kaess, M.: Windowed bundle adjustment framework for unsupervised learning of monocular depth estimation with u-net extension and clip loss. IEEE Robotics and Automation Letters **5**(2), 3283–3290 (2020)
31. Zou, Y., Luo, Z., Huang, J.B.: DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency. In: ECCV (2018)