Photo-SLAM: Real-time Simultaneous Localization and Photorealistic Mapping for Monocular, Stereo, and RGB-D Cameras
- Supplementary Material -

Huajian Huang[1] Longwei Li[2] Hui Cheng[2] Sai-Kit Yeung[1]
[1]The Hong Kong University of Science and Technology [2]Sun Yat-sen University
hhuangbg@connect.ust.hk, lilw23@mail2.sysu.edu.cn, chengh9@mail.sysu.edu.cn, saikit@ust.hk

Figure 1. The proposed system has real-time performance on embedded platforms, such as Jetson AGX Orin Developer Kit.

Photo-SLAM is a novel system for simultaneous localization and photorealistic mapping, which can even run on embedded platforms at real-time speed, as demonstrated in Fig. 1. In this supplementary, we provide additional results regarding localization and mapping performance.

1. Localication

Stability. As online systems, SLAMs are required to process the incoming frames and estimate current camera poses in time. Therefore, tracking stability regarding latency and the average processing time is an important factor in evaluating system performance in addition to pose estimation accuracy. As reported in Table 1 of the main paper, Photo-SLAM is capable of processing more than 40 frames per second with accurate pose estimation. The average tracking speed is about six times faster than ESLAM [1] and three times faster than Co-SLAM [3]. Here, we provide additional analysis on tracking stability while an example plotted in Fig. 2.

Although the average tracking time of Go-SLAM [4] is less than Co-SLAM and ESLAM, the processing latency

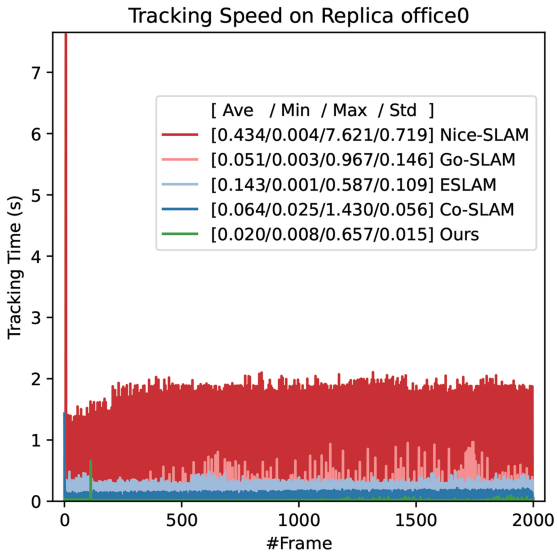


Figure 2. Tracking speed comparisons on scene office0 using an RGB-D camera. The vertical axis denotes the processing time of each frame while the horizontal axis denotes the frame number. [Ave/Min/Max/Std] represent the average, minimum, and maximum tracking time and its standard deviation respectively.

is high due to frequently conducting expensive global optimization. As shown in Fig. 2, Go-SLAM often takes about 1 second to process the frame and estimate the pose. Moreover, both Nice-SLAM [5] and Co-SLAM need a longer time to accurately initialize the tracking. Obviously, our system can rapidly and stably process the incoming frames, having minimum average tracking time and standard deviation. The peak processing time of our system occurs when loop closure is detected for correcting pose estimation drift.

Accuracy. Some qualitative tracking results of Photo-SLAM are demonstrated in Fig. 3.

(a) Replica office0

(b) Replica room0

(c) TUM fr3-office

Figure 3. Trajectory in the reconstructed map. Green points denote ground truth trajectory while red denotes the estimated trajectory of Photo-SLAM.

method	PSNR ↑	SSIM ↑	LPIPS ↓	Rendering FPS ↑	Model Size (MB)
1) 3DG	27.844	0.861	0.213	745.480	36.141
2) 3DG	34.555	0.942	0.065	483.904	144.196
3) 3DG	37.055	0.962	0.032	448.109	219.470
Ours using Mono	33.302	0.926	0.078	911.262	31.419
Ours using RGB-D	34.958	0.942	0.059	1084.017	35.211

Table 1. Comparison of mapping performance between 3D Gaussian splatting (3DG) and our system Photo-SLAM with different settings on the Replica dataset.

On TUM Dataset			Resources		
Scene	Cam	Method	Tracking FPS ↑	Rendering FPS ↑	Model Size (MB)
fr1-desk	Mono	**Ours (Jetson)**	28.267	340.507	4.610
		Ours (Laptop)	28.330	1105.062	7.421
		Ours	57.781	2016.690	10.027
	RGB-D	**Ours (Jetson)**	27.970	380.622	5.743
		Ours (Laptop)	28.930	1061.040	8.432
		Ours	58.378	2083.896	9.963
fr2-xyz	Mono	**Ours (Jetson)**	24.005	169.321	14.286
		Ours (Laptop)	24.922	619.554	16.102
		Ours	58.241	1405.797	20.380
	RGB-D	**Ours (Jetson)**	21.032	274.718	6.319
		Ours (Laptop)	22.665	701.590	13.850
		Ours	52.904	1790.120	21.399
fr3-office	Mono	**Ours (Jetson)**	36.700	291.398	10.669
		Ours (Laptop)	38.929	824.658	16.249
		Ours	81.575	1522.120	19.211
	RGB-D	**Ours (Jetson)**	18.039	291.907	12.726
		Ours (Laptop)	19.636	764.342	15.349
		Ours	43.650	1540.757	17.009

Table 2. Additional results of Photo-SLAM with different platforms on the TUM dataset.

2. Discussion

Online Mapping vs Offline Mapping. For online mapping, the mapping process occurs simultaneously with the localization process. Therefore it requires continuous and prompt updates with each new observation as the robot or camera moves and observes its surroundings. In general, online photorealistic mapping is more challenging than offline photorealistic mapping, since it is crucial to balance the trade-off between computational efficiency and rendering quality. As mentioned in the main paper, we proposed a geometry-based densification strategy and a Gaussian-Pyramid-based (GP) learning method to achieve high-quality online mapping. To further support this statement, we compared the photorealistic mapping performance between our Photo-SLAM and 3D Gaussian splatting (3DG) [2]. 3DG is the SOTA offline method which takes a set of images with known poses and a sparse point cloud as input to learn a radiance field for view synthesis. During the experiments, 3DG used the keyframe poses estimated by Photo-SLAM and performed training for the same duration as Photo-SLAM. The required point cloud input is initialized in three different ways: 1) randomly initializing 100 points; 2) randomly initializing 10,000 points; and 3) initializing from the hyper primitives map of Photo-SLAM. The results are reported in Table 1. Without inputting fine-grained point clouds, 3DG needs more time for optimization such that the rendering quality decreases. In addition, to enhance rendering quality, 3DG tends to densify point clouds leading to larger model size and slower rendering speed. Whether using monocular cameras or RGB-D cameras, Photo-SLAM consistently delivers compelling rendering quality and faster rendering speeds, owing to the effectiveness of the proposed algorithms.

On EuRoC Dataset		Resources		
Scene	Method	Tracking FPS ↑	Rendering FPS ↑	Model Size (MB)
MH-01	**Ours (Jetson)**	21.359	93.762	43.385
	Ours (Laptop)	25.019	316.403	89.700
	Ours	44.977	613.958	123.528
MH-02	**Ours (Jetson)**	22.355	101.021	36.263
	Ours (Laptop)	26.189	332.174	81.569
	Ours	46.556	675.508	113.116
V1-01	**Ours (Jetson)**	21.332	106.008	28.444
	Ours (Laptop)	25.403	367.903	55.263
	Ours	44.763	835.119	74.457
V2-01	**Ours (Jetson)**	23.872	99.988	27.840
	Ours (Laptop)	27.556	307.025	62.588
	Ours	48.911	595.234	82.600

Table 3. Additional results of Photo-SLAM with different platforms on the EuRoC MAV stereo dataset.

3. More Results

The results of each scene of the replica dataset are detailed in Table 4 Additional qualitative results on the TUM dataset are demonstrated on Fig. 4 and Fig. 5, while Fig. 6 illustrates qualitative results of Photo-SLAM on EuReC Stereo Dataset.

References

[1] Mohammad Mahdi Johari, Camilla Carta, and François Fleuret. Eslam: Efficient dense slam system based on hybrid representation of signed distance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17408–17419, 2023.

[2] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4): 1–14, 2023.

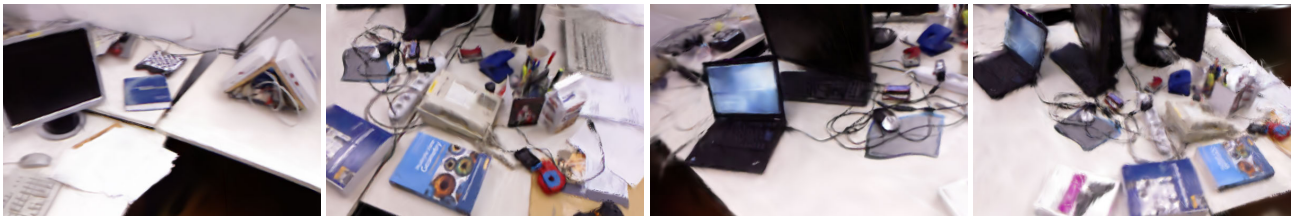
[3] Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13293–13302, 2023.

[4] Youmin Zhang, Fabio Tosi, Stefano Mattoccia, and Matteo Poggi. Go-slam: Global optimization for consistent 3d instant reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3727–3737, 2023.

[5] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022.

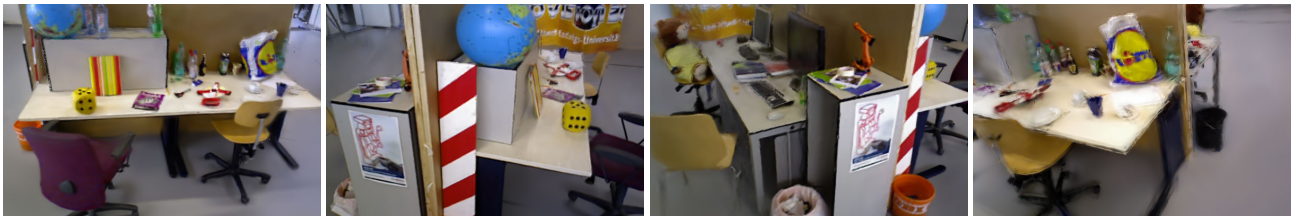
On Replica Dataset			Localization		Mapping			Resources		
Scene	Cam	Method	Trajectory ($RMSE_{cm}$) ↓	Rotation (RMSE) ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Tracking FPS ↑	Rendering FPS ↑	Model Size (MB)
office0	Mono	**Ours (Jetson)**	0.467	0.00334	34.415	0.940	0.949	18.328	123.551	17.703
		Ours (Laptop)	0.587	0.00343	36.227	0.954	0.071	20.422	413.645	21.703
		Ours	0.575	0.00369	36.989	0.955	0.061	42.487	930.598	24.975
	RGB-D	**Ours (Jetson)**	0.499	0.00356	35.447	0.949	0.086	19.076	154.087	18.281
		Ours (Laptop)	0.519	0.00317	38.219	0.965	0.053	22.446	497.917	18.826
		Ours	0.522	0.00307	38.477	0.964	0.050	48.588	1447.887	19.740
office1	Mono	**Ours (Jetson)**	5.586	0.31140	32.382	0.904	0.113	18.312	80.048	22.904
		Ours (Laptop)	0.379	0.00463	37.970	0.954	0.060	19.968	322.160	21.224
		Ours	0.315	0.00383	37.592	0.950	0.062	42.296	857.324	26.982
	RGB-D	**Ours (Jetson)**	0.402	0.00517	37.510	0.953	0.065	19.194	118.584	20.656
		Ours (Laptop)	0.440	0.00543	39.109	0.962	0.049	22.349	496.644	19.548
		Ours	0.436	0.00477	39.089	0.961	0.047	47.333	1263.343	21.193
office2	Mono	**Ours (Jetson)**	1.402	0.01452	28.083	0.900	0.131	17.502	90.181	19.704
		Ours (Laptop)	2.087	0.02154	31.202	0.927	0.098	18.975	343.662	27.332
		Ours	5.031	0.04696	31.794	0.929	0.091	39.604	930.777	31.558
	RGB-D	**Ours (Jetson)**	1.209	0.00964	29.755	0.919	0.110	17.860	124.420	27.560
		Ours (Laptop)	1.188	0.00972	32.720	0.940	0.080	21.507	425.452	31.711
		Ours	1.276	0.01094	33.034	0.938	0.077	44.062	904.249	34.065
office3	Mono	**Ours (Jetson)**	0.429	0.00232	28.058	0.886	0.132	17.881	96.872	15.505
		Ours (Laptop)	0.409	0.00239	32.012	0.924	0.090	19.518	368.530	20.475
		Ours	0.472	0.00227	31.622	0.920	0.086	40.870	1131.957	26.653
	RGB-D	**Ours (Jetson)**	0.718	0.00222	30.954	0.917	0.103	17.889	120.118	20.270
		Ours (Laptop)	0.747	0.00233	33.594	0.939	0.072	20.051	388.624	23.617
		Ours	0.782	0.00233	33.789	0.938	0.066	40.603	1125.175	25.226
office4	Mono	**Ours (Jetson)**	0.579	0.00305	30.399	0.921	0.109	18.755	102.949	15.201
		Ours (Laptop)	0.616	0.00279	33.656	0.940	0.078	20.311	375.033	21.444
		Ours	0.583	0.00272	34.168	0.941	0.072	42.262	849.305	26.154
	RGB-D	**Ours (Jetson)**	0.661	0.00367	32.219	0.931	0.091	17.107	92.237	32.405
		Ours (Laptop)	0.629	0.00446	35.534	0.951	0.059	19.361	333.874	32.270
		Ours	0.582	0.00423	36.020	0.952	0.054	39.870	1061.749	35.421
room0	Mono	**Ours (Jetson)**	0.369	0.00321	26.423	0.787	0.221	17.987	87.246	17.121
		Ours (Laptop)	0.349	0.00294	29.899	0.868	0.125	19.521	332.127	33.151
		Ours	0.345	0.00299	29.772	0.871	0.106	41.020	754.729	44.333
	RGB-D	**Ours (Jetson)**	0.514	0.00265	27.867	0.833	0.165	17.424	104.248	31.196
		Ours (Laptop)	0.521	0.00257	31.288	0.914	0.075	19.119	322.585	52.266
		Ours	0.541	0.00270	30.716	0.899	0.075	39.825	897.870	55.397
room1	Mono	**Ours (Jetson)**	0.803	0.00670	27.076	0.841	0.177	19.834	99.038	19.699
		Ours (Laptop)	1.046	0.00868	30.459	0.902	0.092	21.580	333.430	32.959
		Ours	1.183	0.00772	31.302	0.910	0.083	44.316	782.326	43.865
	RGB-D	**Ours (Jetson)**	0.381	0.00299	30.191	0.895	0.108	18.881	121.986	29.503
		Ours (Laptop)	0.399	0.00277	33.071	0.931	0.062	21.782	367.455	43.568
		Ours	0.394	0.00320	33.511	0.934	0.057	43.352	1018.111	49.617
room2	Mono	**Ours (Jetson)**	0.241	0.00280	27.432	0.889	0.138	17.918	80.568	16.303
		Ours (Laptop)	0.235	0.00263	32.970	0.935	0.075	19.499	339.442	22.790
		Ours	0.225	0.00258	33.181	0.934	0.067	40.313	1053.078	26.834
	RGB-D	**Ours (Jetson)**	0.260	0.00257	31.883	0.928	0.078	15.982	95.480	33.592
		Ours (Laptop)	0.275	0.00250	35.295	0.954	0.045	18.158	336.103	38.307
		Ours	0.305	0.00257	35.028	0.951	0.043	36.244	953.755	41.032

Table 4. Detailed results of Photo-SLAM with different platforms on the Replica dataset.



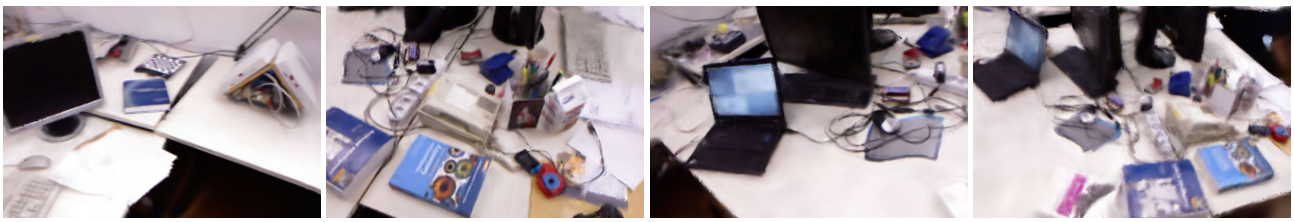
(a) fr1-desk (Mono)

(b) fr2-xyz (Mono)



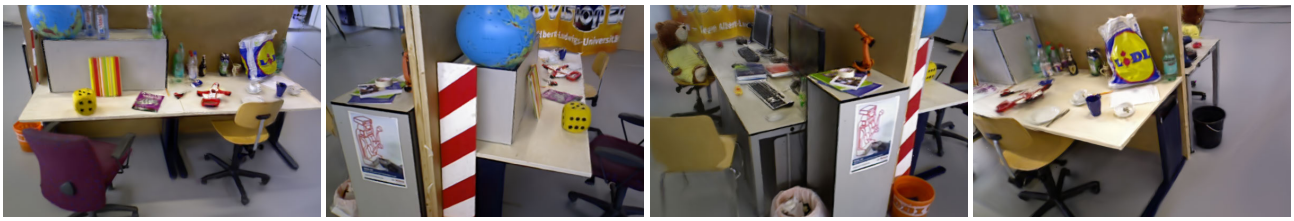
(c) fr3-office (Mono)

Figure 4. Qualitative results of Photo-SLAM on TUM using monocular cameras.



(a) fr1-desk (RGB-D)

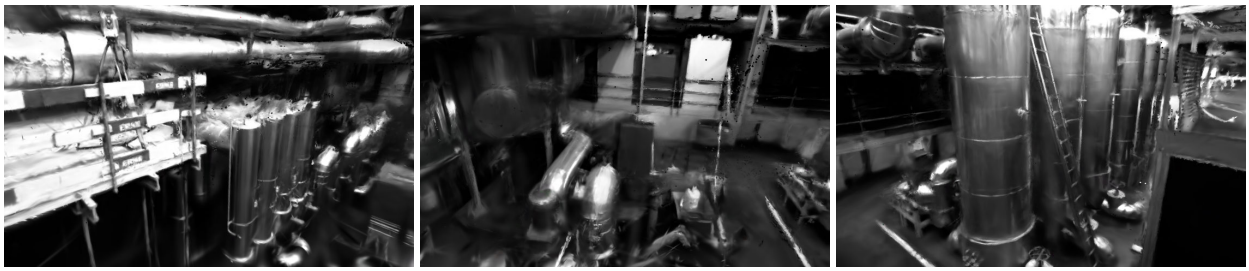
(b) fr2-xyz (RGB-D)



(c) fr3-office (RGB-D)

Figure 5. Qualitative results of Photo-SLAM on TUM using RGB-D cameras.

(a) EuRoC MH-01



(b) EuRoC MH-02

(c) EuRoC V1-01

(d) EuRoC V2-01

Figure 6. Qualitative results of Photo-SLAM on stereo EuRoC.