-Supplementary Material-
iS-MAP: Neural Implicit Mapping and Positioning for Structural Environments

Haocheng Wang[1], Yanlong Cao[1*], Yejun Shou[1], Lingfeng Shen[1], Xiaoyao Wei[1], Zhijie Xu[2], and Kai Ren[1]

[1] College of Mechanical Engineering, Zhejiang University, Hangzhou, 310027, China
00whcl@zju.edu.cn sdcaoyl@zju.edu.cn
[2] Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China
Zhijie.Xu@xjtlu.edu.cn

1 Per-Scene Breakdown of the Results

In this section, we breakdown the quantitative analysis of Tab. 1 in the main text into a per-scene analysis. Tab. 1 shows the per-scene quantitative evaluation of our method in comparison with otherNeRF-based SLAM methods on the Replica dataset [4].

2 More ablation of scene representation

We conducted detailed ablation experiments on scene representation and presented the results in Tab. 2. Methods without hash grids showed a slight advantage in Comp.Ratio, but other metrics exhibited a decline. Unlike the baseline [2] using coarse and fine high-dimensional feature planes, the feature plans used in our method have multiple resolutions with fewer dimensions. While this enhances detail perception, it struggles to represent areas with complex variations. Since Replica [4] is a synthetic dataset with few noise. In most cases, only using the lower-dimensional feature plane can encode the space well, making the gain from the hash grid relatively low(such as reduced Comp.Ratio in Replica [4]). In contrast, ScanNet [1] is a challenging dataset collected by handheld devices in real-world, characterized by more noise and lower integrity. The hash grid can provide a supplementary representation for these regions with less computational overhead, thereby enhancing tracking and mapping.

3 Qualitative ablation for structural consistency

Structural constraint is one of the key contributions of this paper. By utilizing prior structural consistency constraints, spatial planes and lines can be better regularized in mapping stage. As shown in the qualitative ablation results of ScanNet [1] in Fig. 1, methods with structural consistency reconstructs a more complete desk and cabinet, reducing the artifacts and floaters.

Table 1: Per-scene comparison of reconstruction accuracy for our method and other NeRF-based SLAM methods. The best results were highlighted in red and the second best results were highlighted in blue.

Methods	Metric	room0	room1	room2	office0	office1	office2	office3	office4	Avg
iMAP [5]	**Depth L1**[cm]↓	5.08	3.44	5.78	3.79	3.76	3.97	5.61	5.71	4.64
	Acc.[cm]↓	4.01	3.04	3.84	3.34	2.10	4.06	4.20	4.34	3.62
	Comp.[cm]↓	5.84	4.40	5.07	3.62	3.62	4.73	5.49	6.65	4.93
	Comp.Ratio[%]↑	78.34	85.85	79.40	83.89	88.45	79.73	73.90	74.77	80.50
	RMSE[cm]↓	3.12	2.54	2.31	1.69	1.03	3.99	4.05	1.93	2.58
NICE-SLAM [9]	**Depth L1**[cm] ↓	1.79	1.33	2.20	1.43	1.58	2.70	2.10	2.06	1.90
	Acc.[cm]↓	2.44	2.10	2.17	1.85	1.56	3.28	3.01	2.54	2.37
	Comp.[cm]↓	2.60	2.19	2.73	1.84	1.82	3.11	3.16	3.61	2.63
	Comp.Ratio[%]↑	91.81	93.56	91.48	94.93	94.11	88.27	87.68	87.23	91.13
	RMSE[cm]↓	1.69	2.04	1.55	0.99	0.90	1.39	3.97	3.08	1.95
Vox-Fusion [8]	**Depth L1**[cm] ↓	1.76	2.52	3.58	3.44	1.77	3.52	1.82	4.84	2.91
	Acc.[cm]↓	**1.77**	**1.51**	2.23	1.63	1.60	**2.02**	**2.33**	**2.02**	**1.88**
	Comp.[cm]↓	2.69	2.31	2.58	1.87	1.66	3.03	2.81	3.51	2.56
	Comp.Ratio[%]↑	92.03	92.47	90.13	93.86	94.40	88.94	89.10	86.53	90.94
	RMSE[cm]↓	1.37	1.90	1.47	1.35	1.76	1.18	1.11	1.64	1.03
Structerf-SLAM [6]	**Depth L1**[cm] ↓	1.70	1.54	2.13	1.47	1.56	2.22	2.21	2.06	1.86
	Acc.[cm]↓	2.33	2.24	2.05	1.81	1.60	3.03	2.94	2.46	2.30
	Comp.[cm]↓	2.60	2.30	2.29	1.88	1.72	2.99	3.19	3.54	2.56
	Comp.Ratio[%]↑	92.16	93.62	92.58	94.89	94.47	89.17	87.32	87.11	91.42
	RMSE[cm]↓	0.68	**0.45**	0.70	0.57	0.50	1.18	0.94	2.01	0.88
Co-SLAM [7]	**Depth L1**[cm] ↓	1.05	0.85	2.37	1.24	1.48	1.86	1.66	1.54	1.51
	Acc.[cm]↓	2.11	1.68	1.99	1.57	**1.31**	2.84	3.06	2.23	2.10
	Comp.[cm]↓	2.02	1.81	1.96	1.56	1.59	2.43	2.72	2.52	2.08
	Comp.Ratio[%]↑	95.26	95.19	93.58	96.09	94.65	91.63	90.72	90.44	93.44
	RMSE[cm]↓	0.65	1.13	1.43	0.55	0.50	0.46	1.40	0.77	0.86
ESLAM [2]	**Depth L1**[cm] ↓	0.73	0.74	1.26	0.71	1.02	0.93	1.03	1.18	0.95
	Acc.[cm]↓	2.15	1.94	1.68	1.61	1.82	2.95	2.55	2.10	2.08
	Comp.[cm]↓	1.79	1.58	1.65	1.45	1.30	1.92	2.20	2.13	1.75
	Comp.Ratio[%]↑	**97.39**	96.50	**96.99**	**98.45**	97.60	95.07	95.05	94.31	96.43
	RMSE[cm]↓	0.71	0.70	0.52	0.57	0.55	0.58	0.72	0.63	0.63
Ours	**Depth L1**[cm]↓	**0.62**	**0.55**	**0.86**	**0.53**	**0.92**	**0.88**	**0.80**	**0.86**	**0.75**
	Acc.[cm]↓	2.23	1.66	**1.67**	**1.50**	1.44	2.47	2.57	2.16	1.96
	Comp.[cm]↓	**1.78**	**1.53**	**1.60**	**1.31**	**1.22**	**1.82**	**2.01**	**2.01**	**1.66**
	Comp.Ratio[%]↑	97.21	**97.04**	96.94	98.16	**97.44**	**95.94**	**95.91**	**94.45**	**96.64**
	RMSE[cm]↓	**0.58**	**0.57**	**0.45**	**0.36**	**0.35**	**0.46**	**0.57**	**0.49**	**0.48**

Table 2: The average results of the scene representation ablation experiments across 8 scenes in Replica and 5 scenes in ScanNet.

	Replica					ScanNet
	Depth L1	Acc.	Comp.	Comp.Ratio	RMSE	RMSE
No Feature Plane	1.78	2.91	2.11	93.45	0.95	6.98
No Hash Grid	0.82	2.11	1.68	**96.67**	0.59	7.02
ours	**0.75**	**1.96**	**1.66**	96.64	**0.48**	**6.57**

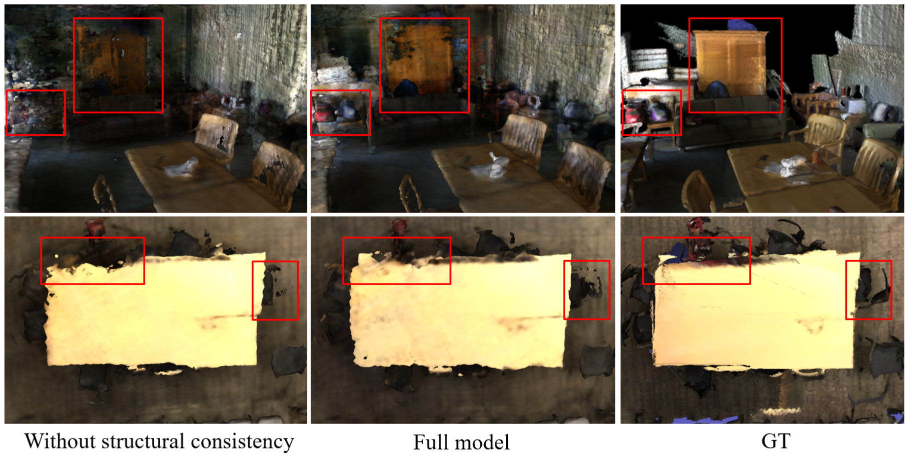

Fig. 1: Qualitative ablation results for structural consistency.

4 Discussion about color loss

Among the metrics in Tab. 4 of the main text, mapping metrics like Acc. and Comp. focus on spatial points accuracy, with color having no direct impact on them. But the absence of color loss reduces tracking performance (RMSE) and makes the reconstruction result entirely colorless. This is consistent with the conclusion of [2, 3, 9]. As shown in Fig. 2, the colorless door and windows are almost indistinguishable, which is unacceptable for visualization.

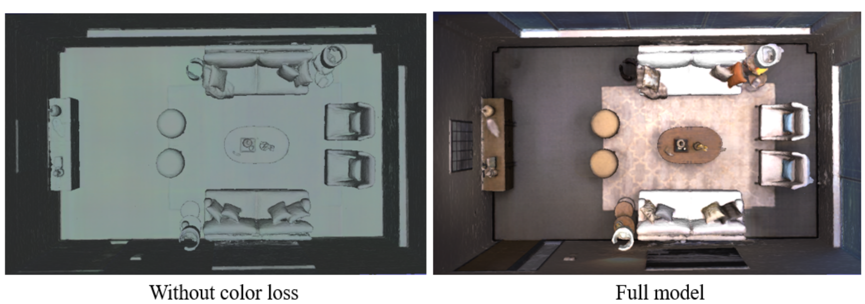

Fig. 2: The reconstruction results of Replica room0 with/without color loss.

References

1. Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor

scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 1

2. 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. 1, 2, 3
3. Erik Sandström, Yue Li, Luc Van Gool, and Martin R Oswald. Point-slam: Dense neural point cloud-based slam. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18433–18444, 2023. 3
4. Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 1
5. Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021. 2
6. Haocheng Wang, Yanlong Cao, Xiaoyao Wei, Yejun Shou, Lingfeng Shen, Zhijie Xu, and Kai Ren. Structerf-slam: Neural implicit representation slam for structural environments. *Computers & Graphics*, page 103893, 2024. 2
7. 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. 2
8. Xingrui Yang, Hai Li, Hongjia Zhai, Yuhang Ming, Yuqian Liu, and Guofeng Zhang. Vox-fusion: Dense tracking and mapping with voxel-based neural implicit representation. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 499–507, 2022. 2
9. 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. 2, 3