GO-SLAM: Global Optimization for Consistent 3D Instant Reconstruction – Supplementary Material

Youmin Zhang Fabio Tosi Stefano Mattoccia Matteo Poggi
Department of Computer Science and Engineering (DISI)
University of Bologna, Italy
{youmin.zhang2, fabio.tosi5, stefano.mattoccia, m.poggi}@unibo.it
https://youmi-zym.github.io/projects/GO-SLAM/

A. SLAM Configuration

A.1. Rendering Network Details.

Architecture. Our rendering networks consist of one SDF network and one color network as shown in Fig. A. For multi-resolution hash encodings, we set the feature dimensions and maximum entries per level to 2 and 2^{19} respectively. Totally, there are 16 levels of different resolutions ranging from 16 to 4096. The size of hidden units in the SDF network and color network are 32 and 64 respectively.

Training. We use AdamW [4] as optimizer. The learning rate is set to 10^{-2} for learnable hash encodings and 10^{-3} for parameters of the SDF network and color network.

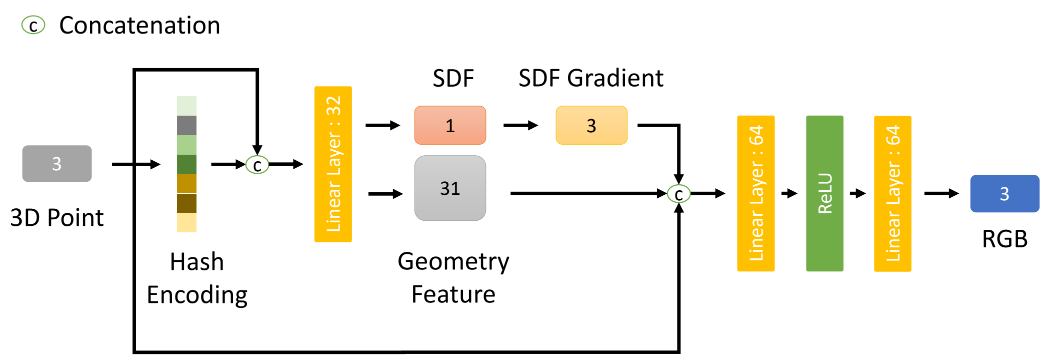

Figure A: **Details of rendering networks.**

B. More Qualitative Results

We provide detailed results on ScanNet [3], EuRoC [1] and Replica [5] datasets on Figs. B to D respectively.

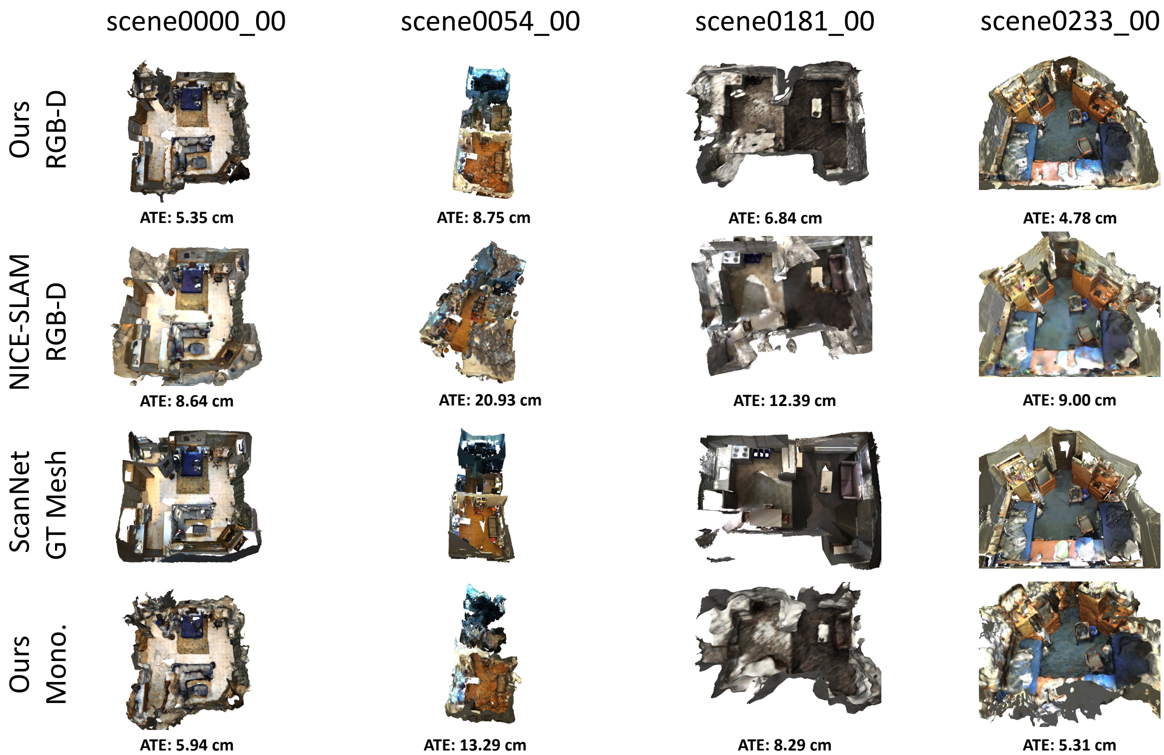

Figure B: **Qualitative results on ScanNet dataset [5].** Our GO-SLAM predicts globally consistent 3D reconstruction with monocular or RGB-D input at real-time speed. While NICE-SLAM [7] suffers from accumulated errors in trajectory and distortion in 3D models.

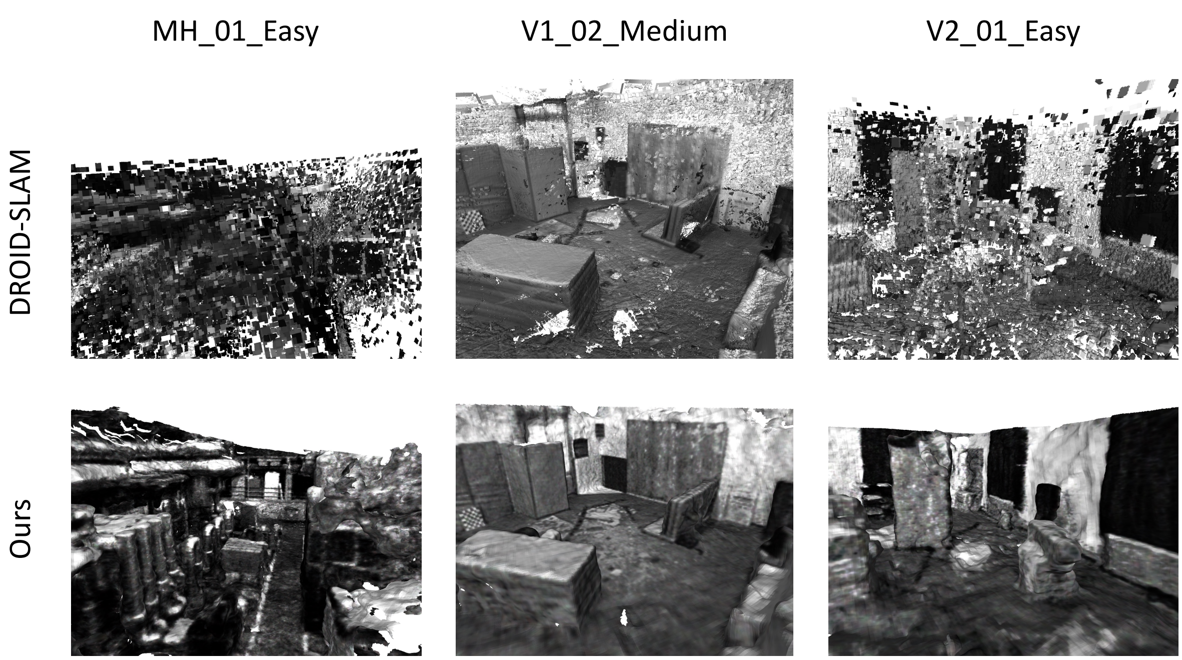

Figure C: **Qualitative results on EuRoC dataset [1].** DROID-SLAM [6] reconstructions obtained with TSDF-Fusion [2]. Compared to DROID-SLAM, our reconstructions are much cleaner and more complete.

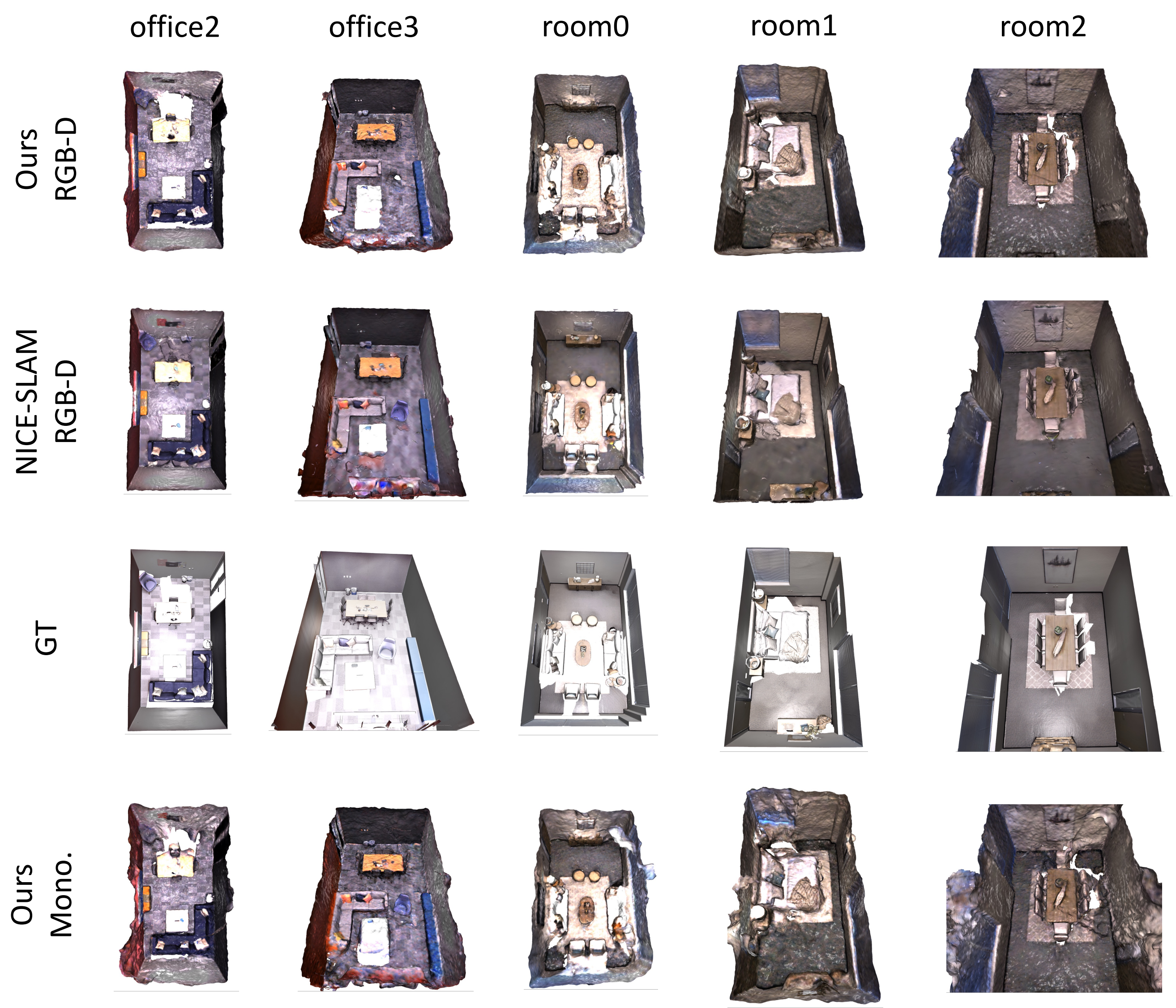

Figure D: **Qualitative results on Replica dataset [5].** Our GO-SLAM predicts dense 3D reconstruction with monocular or RGB-D input at real-time speed. While NICE-SLAM [7], which can only work with depth input, takes hours ($\ll$ 1 FPS) to get the reconstruction.

References

[1] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *IJRR*, 35(10):1157–1163, 2016. 1, 2

[2] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH*, pages 303–312, 1996. 2

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 1

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[5] 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, 2, 3

[6] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *NeurIPS*, 34:16558–16569, 2021. 2

[7] 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 *CVPR*, pages 12786–12796, 2022. 2, 3