NeSLAM: Neural Implicit Mapping and Self-Supervised Feature Tracking With Depth Completion and Denoising (Supplementary Material)

Tianchen Deng, Yanbo Wang, Hongle Xie, Jingchuan Wang, *Senior Member, IEEE*, Danwei Wang, *Life Fellow, IEEE*, Weidong Chen, *Member, IEEE*

I. Overview

In this supplementary material, we provide detailed network parameters and implementation details in Sec.II. Sec.III contains the extensive experiments on various datasets.

II. Implementation Details

Depth Denoising Network For our depth completion and denoising network, we use a ResNet-18 [1] encoder and add a branch for uncertainty estimation. Both branches use a NLSPN module [2] to increase the performance with sparse depth input. The depth completion network is pre-trained at a lower resolution 240×320 for ScanNet datasets. We use a Adam [3] optimizer ($\beta = (0.9, 0.999)$) with a learning rate of 0.0001 and a batch size of 8. The depth denoising network is also fine-tuned with depth loss with the system operation.

Hierarchical Scene Representation Inspired by VolSDF [4], we change the occupancy with Signed Distance Field (SDF) value which greatly improve the ability of geometry representation. The feature dimension is 64 and 5 layers for geometry and 2 layers for color decoders. The coarse level and middle level feature grid is randomly initialized in all experiments. Empirically, the random initialization gives slightly better convergence compared to the fixed feature vector initialization. For the fine level feature grid, it is initialized to ensure the output of the fine level decoder as zero, since it is added in a residual manner onto the occupancy predicted from the mid level feature grid. We use Adam optimizer for scene geometry optimization.

Nerf-Based Self-Supervised Keypoint Detection We use the encoder provided in [5]. This encoder has a VGG-like architecture that has eight 3×3 convolution layers sized 64-64-64-64-128-128-128-128. There is a 2x2 max pool layer for every layer. The keypoint detection decoder head has a single 3x3 convolutional layer of 256 units followed by a 1x1 convolution layer with 65 units and 256 units for the keypoint detector. All convolution layers in the network are followed by ReLU non-linear activation and BatchNorm normalization. We use Adam optimizer for keypoint detection and camera tracking. The learning rate for keypoint detection is 1×10^{-3}. The learning rate for tracking on Replica, ScanNet, TUM RGB-D is 1×10^{-3}, 5×10^{-4}, 1×10^{-2}. For the re-implementation of iMAP: iMAP*, we use the same settings from NICE-SLAM [6].

TABLE I
Frame loss experiments on Replica and ScanNet datasets. Replica* and ScanNet* represent the processed RGB-D sequences.

Methods	Metrics	Replica*	ScanNet*
iMAP*	RMSE[cm]	2.973	F
	Median[cm]	2.762	F
	Mean[cm]	2.593	F
NICE-SLAM	RMSE[cm]	2.586	F
	Median[cm]	2.397	F
	Mean[cm]	2.274	F
NeSLAM	RMSE[cm]	**1.762**	**13.634**
	Median[cm]	**1.534**	**12.427**
	Mean[cm]	**1.489**	**12.046**

III. Additional Experiments

In this section, we provide more experiments of our method on different datasets.

A. Qualitative Results

Due to the space limitation, we put the additional qualitative results into supplementary materials. In Fig. 1, we present the novel view synthesis results of TUM RGB-D datasets fr1/desk. It can be seen that our method achieve better performance for view synthesis.

B. Frame Loss Robustness

We simulate the frame loss on Replica [7] and ScanNet [8] datasets. We randomly remove some frames in these two datasets. We skip one or two frames for every interval, such as ID 20 to 22, ID 110 to 113. In Table I, it can be seen that iMAP and NICE-SLAM struggles to recover camera pose and scene geometry. F means this method fails for camera tracking. Their camera tracking accuracy drop dramatically. In contrast, our method perform better robustness compared with NICE-SLAM and iMAP. This is due to the use of keypoint detection network and patch-wise color and depth loss which improves the robustness of our system.

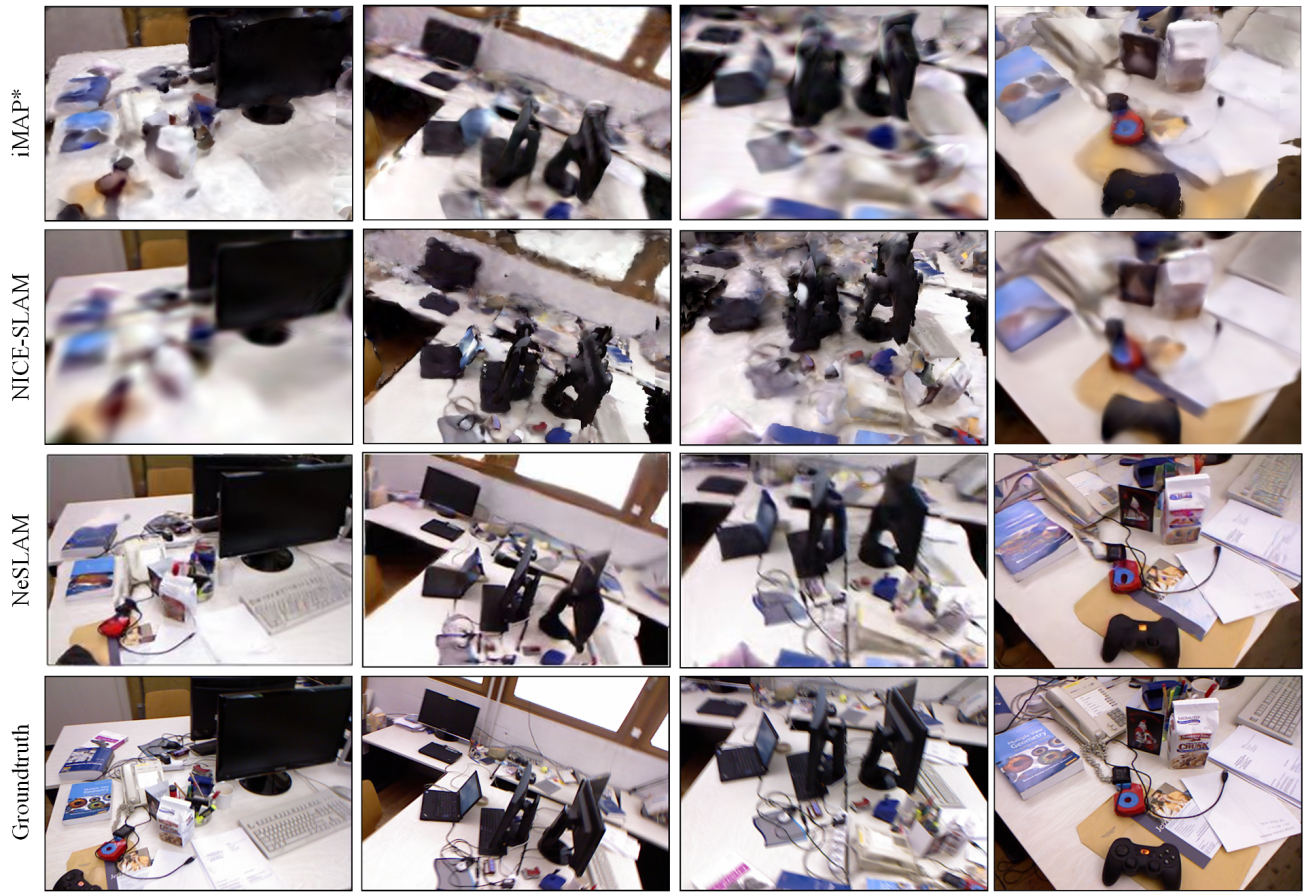

Fig. 1. Qualitative Results on TUM RGB-D dataset [9]. From left to right, we show different view synthesis results on fr1/desk sequence.

TABLE II
DETAILED TIME CONSUMPTION OF OUR SYSTEM.

	Tracking[ms]	Mapping[ms]
iMAP	101	448
NICE-SLAM	50	145
Ours	**44**	**130**

C. Real-Time Performance

In this section, we provide the real-time performance of our system. We use a highly efficient multi-process implementation for the parallel tracking and mapping. It takes 44ms for tracking a new frame, and 147ms for mapping a new frame. Our method demonstrates superior speed compared to NICE-SLAM and significant improvements over iMAP.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.
[2] J. Park, K. Joo, Z. Hu, C.-K. Liu, and I. So Kweon, "Non-local spatial propagation network for depth completion," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 120–136.
[3] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, 2015.
[4] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman, "Volume rendering of neural implicit surfaces," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 4805–4815.
[5] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2018.
[6] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, "Nice-slam: Neural implicit scalable encoding for slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 12 786–12 796.
[7] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma *et al.*, "The replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.
[8] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Niessner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.
[9] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 573–580.