SNI-SLAM: Semantic Neural Implicit SLAM

(Supplementary Material)

Siting Zhu[1]*, Guangming Wang[2]*, Hermann Blum[3], Jiuming Liu[1],
Liang Song[4], Marc Pollefeys[3], Hesheng Wang[1]†
[1] Department of Automation, Shanghai Jiao Tong University [2] University of Cambridge
[3] ETH Zürich [4] China University of Mining and Technology, China
{zhusiting,liujiuming,wanghesheng}@sjtu.edu.cn gw462@cam.ac.uk
{hermann.blum,marc.pollefeys}@inf.ethz.ch TS21060167P31@cumt.edu.cn

1. Overview

In the supplementary material, the chapters are briefly described as follows:

- we present a detailed experimental setup including baseline introduction, implementation details, and evaluation metrics in Sec. 2.
- Additional experimental results are given in Sec. 3 to demonstrate the excellent performance of our method.
- We show visualization results in Sec. 4 on different scenes.
- We give a video demo of real-time mapping and tracking in Sec. 5.

2. Experimental Setup

2.1. Baselines

To the best of our knowledge, NIDS-SLAM [3] is the only existing semantic NeRF-SLAM method. Therefore, we use it as the baseline for comparing the accuracy of semantic NeRF-based SLAM. However, NIDS-SLAM [3] does not evaluate mesh reconstruction accuracy, so we only compare the metrics of the semantic segmentation accuracy with this method. To provide a more comprehensive baseline for comparison, we also consider Semantic-NeRF [13], an offline-trained approach that focuses on high-fidelity semantic 3D reconstruction through hours of extensive optimization. For SLAM accuracy, we compare our method with state-of-the-art NeRF-based SLAM methods [4, 7, 10–12, 14]. Since iMAP [10] is not open source, we use iMAP* [10] in our experiment, which is the reimplementation of iMAP.

*Equal Contribution.
†Corresponding Author.

2.2. Implementation Details

Hyperparameters. For geometry representation, we adopt the coarse feature planes with a resolution of 24 cm and the fine feature planes with a resolution of 6 cm. For semantic and appearance representation, we employ the coarse feature planes with a resolution of 24 cm and the fine feature planes with a resolution of 3 cm. We use 16-channel feature vectors to represent semantic, geometry, and appearance features for both coarse and fine feature planes, resulting in 32-channel concatenated features input for the decoder.

For rendering, we first sample N_{strat} points for each ray by stratified sampling. We then additionally sample N_{imp} points near surfaces. For pixels with ground truth depths, the N_{imp} additional points are uniformly sampled within the truncation distance with respect to the depth measurement. For Replica dataset [8], we set $N_{strat} = 32$ and $N_{imp} = 8$. We conduct 15 optimization iterations for the mapping process and 8 optimization iterations for the tracking process. We sample 4000 pixels for mapping optimization and 2000 pixels for tracking optimization in each iteration. Since ScanNet dataset [2] scenes are at a larger scale and more complicated, we set $N_{strat} = 48$ and $N_{imp} = 8$. We perform 40 optimization iterations for both the mapping and tracking process.

We use a window of 5 keyframes for jointly optimizing scene representation, the MLP network, and the camera poses of the selected keyframes. The weighting coefficients of each loss are $\lambda_{fs} = 5$, $\lambda_s = 0.1$, $\lambda_f = 5$, $\lambda_d = 0.1$, $\lambda_c = 5$. We use a learning rate of 0.005 for feature planes, 0.001 for the decoder, 0.003 for E_θ, H_θ, and F_θ. For camera pose optimization, we use a learning rate of 0.001 in Replica dataset [8]. In ScanNet [2] and TUM RGBD [9] dataset, we use a learning rate of 0.0005 for camera translation optimization and 0.0025 for camera rotation optimiza-

tion.

Network details. For the decoder design D_θ, the input geometry features are processed through a two-layer MLP with 16 channels in the hidden layer. This MLP outputs 17-channel vectors, where values of the first channel are used as the output for SDF values. The remaining 16 channels are concatenated with the input semantic features and input appearance features. The concatenated feature vectors are then passed through another two-layer MLP with a hidden size of 16, which outputs the RGB values. We use ReLU activation function for this hidden layer. Tanh and Sigmoid are respectively used for the output layers of SDF and color values. The input semantic features are processed through a three-layer MLP with 256 channels in the hidden layer to output semantic values.

For cross-attention based feature fusion, the geometry MLP E_θ is a three-layer MLP while the appearance MLP H_θ and the fusion MLP F_θ are both two-layer MLPs. The hidden layers of these MLPs have 16 channels. We use ReLU activation function for the hidden layer of H_θ.

2.3. Evaluation Metrics

For mesh reconstruction evaluation, we use *Depth L1 (cm)*, *Accuracy (cm)*, *Completion (cm)*, and *Completion ratio (%)* with a threshold of 5 cm. We remove unobserved areas outside of any camera radius, as well as additional mesh culling to remove noise points following Co-SLAM [11].

Depth L1 (cm): the average absolute error between ground truth depth and reconstructed depth. The depth values are generated by randomly sampling 1000 views from both the reconstructed meshes and the ground truth meshes.

Accuracy (cm): the average distance from sampled points on the reconstructed mesh to their nearest ground truth points.

Completion (cm): the average distance from sampled points on the ground truth mesh to their nearest points on the reconstructed mesh.

Completion ratio (%): the percentage of points in the reconstructed mesh with *Completion* under 5 cm.

For localization accuracy evaluation, we use ATE [9]. including *RMSE (cm)* and *Mean (cm)* metrics. Semantic segmentation is evaluated with respect to mIoU (%) and per-pixel accuracy (%) [6].

3. Additional Experiments

3.1. More Ablation Studies

The number of feature channels. We conduct an ablation study on the effect of using different numbers of feature channels. Fig. 1 demonstrates that as the number of feature channels increases, the semantic rendering results progressively approach the ground truth labels. This can be attributed to the enhanced ability of high-dimensional features to capture and represent more descriptive and diverse information of the environment. However, the increase in the number of feature channels leads to higher parameter numbers and longer computation time, as shown in Tab. 1. Moreover, using 32-channel features for scene representation with only 15 iterations in mapping and 8 iterations in tracking leads to degraded mapping and tracking performance, as the 32-channel feature planes cannot be sufficiently optimized. Based on the above results, it can be concluded that using 16-channel features in scene representation is a trade-off between the accuracy and runtime.

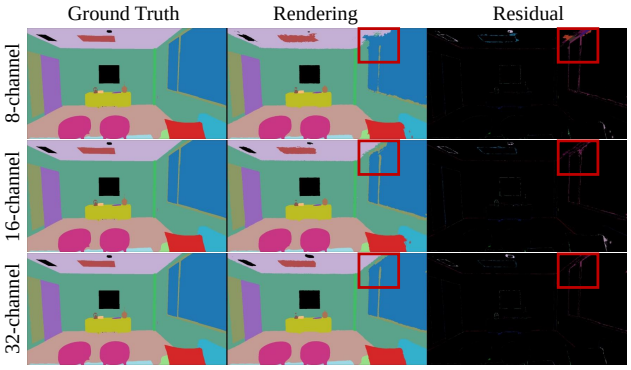

Figure 1. Quantitative analysis on the number of feature channels. With the increase in the number of feature channels, the semantic rendering results achieve higher accuracy.

Channels	Reconstruction Acc. (cm) ↓	Localization RMSE (cm) ↓	Semantic mIoU (%) ↑	#param. ↓	Runtime Map. (ms) ↓	Runtime Track. (ms) ↓
Ours-8	2.21	0.65	82.35	**3.19**	**23.6 × 15**	**7.2 × 8**
Ours-16	**2.09**	**0.50**	84.50	6.23	26.9 × 15	7.8 × 8
Ours-32	2.32	1.64	**84.65**	12.37	34.6 × 15	8.6 × 8

Table 1. Results of using different feature channel numbers. Mapping and tracking runtime is reported in $ms/iter \times iter$ format. With the increase of channel numbers, the numbers of parameter, mapping and tracking time also increase, as the network requires more time to optimize higher-dimensional features.

Effect of noisy semantic results for supervision. Fig. 2 illustrates the robustness of our method when employing noisy semantic supervision signals with various noise ratios. Our approach achieves relatively accurate semantic rendering results even when using noisy semantic supervision with noise ratios up to 90% in both Replica [8] and ScanNet [2] datasets. This robustness is attributable to our feature collaboration strategy, which effectively fuses geometric and appearance information. Specifically, geometric features offer information about the shapes and spatial relationships of objects. Such information can mitigate the effects of inconsistencies caused by noisy semantic labels, primarily because geometric attributes typically maintain their stability despite the presence of semantically noisy labels. Moreover, appearance features provide detailed surface characteristics of objects, such as texture and color,

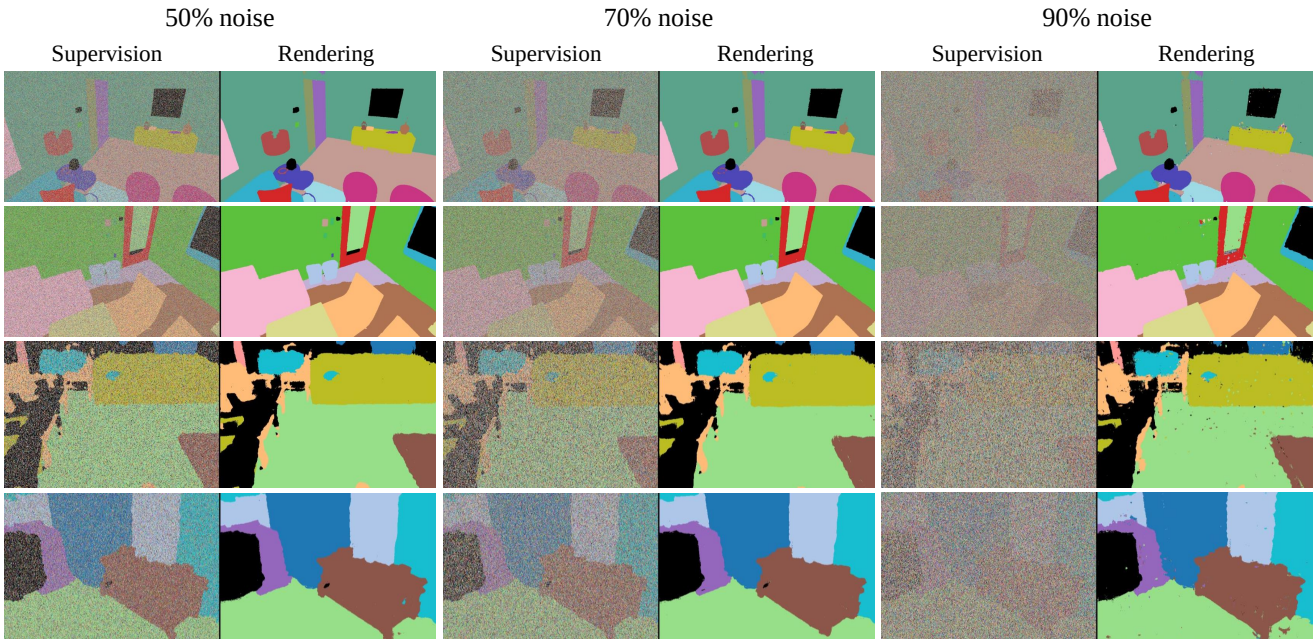

Figure 2. Semantic rendering results with noisy semantic labels for supervision on both Replica [8] and ScanNet [2] datasets.

Name	Reconstruction [cm]			Localization [cm]	
	Acc. ↓	Comp. ↓	Comp.Ratio(%) ↑	RMSE ↓	Mean ↓
w/o semantics	1.83	1.78	96.07	0.58	0.52
w/ semantics	**1.66**	**1.56**	**96.74**	**0.55**	**0.48**

Table 2. Ablation study of adding semantic information. Semantic information can enhance the expression of geometry and appearance, achieving higher accuracy in mapping and tracking.

which are crucial for differentiating similar objects with distinct identities even when their semantic labels may be incorrect. Through the effective collaboration of different features, a more reliable and comprehensive understanding of the scene is achieved. Therefore, we are capable of reducing the negative effects of high-noise labels in the semantic supervision.

Effect of whether to use semantic information. Tab. 2 demonstrates that semantic information can enhance the expression of geometric and appearance information, leading to higher accuracy in mesh reconstruction and localization. Specifically, semantic information provides a richer and more detailed context for the interpretation of geometric and appearance information. For example, knowing the semantic segmentation results of a set of geometric data corresponding to a chair can guide the model to reconstruct and locate the object more accurately and realistically. Therefore, it can be concluded that the integration of semantic information offers a more comprehensive understanding of the environment.

Effect of whether to add Bundle Adjustment (BA). As shown in Tab. 3, we incorporate BA into our SNI-SLAM following BARF [5] and L2G-NeRF [1]. Such integration only shows a slight improvement in SLAM accuracy because current BA NeRFs lack adequate constraints for semantic SLAM optimization.

Name	Acc.[cm]↓	Comp.[cm]↓	Comp.Ratio(%)↑	RMSE[cm]↓	mIoU(%)↑
w/o BA	2.520	**1.873**	95.201	0.320	85.91
w/ BA	**2.517**	1.876	**95.350**	**0.312**	**85.95**

Table 3. Experiment of whether to use BA on Replica [8].

Ablation of our innovations across different scenes. We conduct ablation experiments across 8 scenes in Tab. 4-7, noticing that each innovation is added from baseline. All innovations achieve large increased accuracy (average 13% improvement). Specifically, cross-attention based feature fusion and feature loss increase semantic accuracy by 23.2%, attributed to mutual reinforcement between different modality features of scene and higher-level supervision for semantic optimization. Moreover, one-way correlation decoder can achieve up to 33% improvement in metrics.

3.2. Semantic Segmentation Results

Tab. 8 shows per-scene semantic segmentation results of both online and offline 3D semantic reconstruction methods. For online methods, NIDS-SLAM [3] and our SNI-SLAM, take RGB-D frames as input and perform real-time mapping and camera pose estimation, which require **several minutes** for semantic scene reconstruction. For the offline

method, Semantic-NeRF [13], takes camera pose and RGB-D frames as input and requires nearly **10 hours** of training to obtain semantic reconstruction results, which is several tens of times the duration required by the online methods. Our method outperforms the online method NIDS-SLAM [3] of all scenes and all metrics. Compared with the offline method Semantic-NeRF [13], our real-time semantic mapping method achieves similar results in pixel accuracy metric.

3.3. Reconstruction and Localization Results

Tab. 9 demonstrates per-scene quantitative evaluation of our method with existing NeRF-based SLAM method in Replica dataset [8]. Our method achieves the best performance in *Depth L1*, *Completion*, *Completion ratio (%)*, ATE RMSE, and ATE Mean metrics across all scenes. In some scenes, our method can improve localization accuracy by up to 52% and reconstruction accuracy by up to 32%, demonstrating excellent performance across scenes. This remarkable improvement is attributed to the thoughtful design of the semantic NeRF-based SLAM framework.

4. Visualization

We demonstrate top-view semantic mapping results in Fig. 3, showing that our method is capable of achieving accurate segmentation results, even for objects with complex shapes, such as flowers, and small objects on the table. This capability is due to the hierarchical semantic representation, which provides a coarse-to-fine level of understanding. This approach initially identifies the overall semantic information of the scene, then gradually refines this understanding to capture intricate details, even for small or complex objects. Such representation allows for a more comprehensive perception of the environment, facilitating accurate segmentation results.

Semantic rendering results are shown in Fig. 4. From the residual, we can observe that our method achieves excellent semantic segmentation accuracy in both Replica [8] and ScanNet [2] datasets. Fig. 5 shows pixel accuracy and mIoU changing curve of semantic rendering result from the first frame to the last frame in real-time semantic mapping and tracking. This graph illustrates the fast convergence speed of SNI-SLAM as well as the high semantic accuracy of real-time mapping, which is attributed to loss construction at the feature level. Such high-level guidance for scene optimization is able to accelerate convergence speed as well as achieving accurate semantic mapping.

Fig. 6 and Fig. 7 demonstrate detailed zoom-in views of Replica dataset [8] and Fig. 8 shows results on ScanNet dataset [2]. While other methods fail to reconstruct details such as chair legs and chair back, our method is capable of complete scene reconstruction. This is attributed to the integration of different modality information, as it enables the complementarity of multi-modal information. Moreover, as shown in Fig. 8, our method can achieve a smoother mesh reconstruction due to fully utilizing the advantages of each modality. The above results indicate that our method is capable of retaining more details and achieving a more complete reconstruction compared to existing NeRF-based SLAM methods. Furthermore, our method can provide smoother, more coherent transitions for high-quality reconstruction.

5. Video Demo

We present a video demo, demo.mp4, on room0 of the Replica dataset [8]. In this video, we demonstrate the entire real-time mapping process, including semantic mapping and RGB mapping, as well as the tracking process, showing excellent performance of our method. From the video, we can view accurate semantic segmentation and RGB mapping results of the scene from top view. In addition, it can be observed that the ground truth trajectory and the estimated trajectory almost completely overlap, demonstrating high localization accuracy. We strongly recommend readers to view our video.

References

[1] Yue Chen, Xingyu Chen, Xuan Wang, Qi Zhang, Yu Guo, Ying Shan, and Fei Wang. Local-to-global registration for bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8264–8273, 2023. 3

[2] 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 conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1, 2, 3, 4, 9, 11

[3] Yasaman Haghighi, Suryansh Kumar, Jean Philippe Thiran,

	Name	room0			room1		
		Acc.[cm]↓	RMSE[cm]↓	mIoU(%)↑	Acc.[cm]↓	RMSE[cm]↓	mIoU(%)↑
(a)	Ours (w/o all innovations in ablation studies, **baseline model)**	2.45	0.70	65.12	2.08	1.13	54.60
(b)	Ours (+ feature loss to (a))	2.20	0.55	71.30	1.96	0.92	60.53
(c)	Ours (+ feature fusion with feature loss to (a))	2.13	0.53	75.31	1.85	0.67	67.29
(d)	Ours (+ decoder design to (a))	2.21	0.53	72.56	1.82	0.89	58.91
(e)	Ours (+ hierarchical semantic representation to (a))	2.30	0.55	80.24	1.92	0.79	75.32
(f)	Ours (full, with all innovations in ablation studies, **best model)**	**2.09**	**0.50**	**87.32**	**1.66**	**0.55**	**86.33**

Table 4. Ablation study of our innovations on **room0** and **room1** of Replica [8].

	Name	room2			office0		
		Acc.[cm]↓	RMSE[cm]↓	mIoU(%)↑	Acc.[cm]↓	RMSE[cm]↓	mIoU(%)↑
(a)	Ours (w/o all innovations in ablation studies, **baseline model)**	1.77	0.55	63.91	1.59	0.83	71.52
(b)	Ours (+ feature loss to (a))	1.73	0.50	72.44	1.50	0.59	74.23
(c)	Ours (+ feature fusion with feature loss to (a))	1.70	0.46	75.38	1.48	0.55	77.74
(d)	Ours (+ decoder design to (a))	1.65	0.47	74.28	1.49	0.56	75.94
(e)	Ours (+ hierarchical semantic representation to (a))	1.71	0.51	82.01	1.51	0.55	84.12
(f)	Ours (full, with all innovations in ablation studies, **best model)**	**1.64**	**0.45**	**85.16**	**1.46**	**0.33**	**86.01**

Table 5. Ablation study of our innovations on **room2** and **office0** of Replica [8].

and Luc Van Gool. Neural implicit dense semantic slam. *arXiv preprint arXiv:2304.14560*, 2023. 1, 3, 4, 6

[4] 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, 7

[5] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021. 3

[6] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[7] 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. 1

[8] 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, 4, 5, 6, 7, 8, 9, 10

[9] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012. 1, 2

[10] 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. 1, 7

[11] 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, 7

[12] 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 *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 499–507. IEEE, 2022. 1, 7

[13] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 1, 4, 6

[14] 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. 1, 7

	Name	office1			office2		
		Acc.[cm]↓	RMSE[cm]↓	mIoU(%)↑	Acc.[cm]↓	RMSE[cm]↓	mIoU(%)↑
(a)	Ours (w/o all innovations in ablation studies, **baseline model)**	1.84	0.89	58.65	2.99	0.49	64.36
(b)	Ours (+ feature loss to (a))	1.65	0.74	64.79	2.74	0.43	67.62
(c)	Ours (+ feature fusion with feature loss to (a))	1.62	0.58	67.35	2.70	0.41	72.01
(d)	Ours (+ decoder design to (a))	1.64	0.73	62.92	2.88	0.46	68.45
(e)	Ours (+ hierarchical semantic representation to (a))	1.80	0.65	68.14	2.79	0.45	75.18
(f)	Ours (full, with all innovations in ablation studies, **best model)**	**1.58**	**0.41**	**78.13**	**2.52**	**0.32**	**85.91**

Table 6. Ablation study of our innovations on **office1** and **office2** of Replica [8].

	Name	office3			office4		
		Acc.[cm]↓	RMSE[cm]↓	mIoU(%)↑	Acc.[cm]↓	RMSE[cm]↓	mIoU(%)↑
(a)	Ours (w/o all innovations in ablation studies, **baseline model)**	2.78	0.72	52.42	2.19	0.65	61.92
(b)	Ours (+ feature loss to (a))	2.57	0.66	58.71	2.13	0.59	66.68
(c)	Ours (+ feature fusion with feature loss to (a))	2.55	0.63	61.39	2.09	0.57	68.51
(d)	Ours (+ decoder design to (a))	2.53	0.63	58.88	2.14	0.63	69.25
(e)	Ours (+ hierarchical semantic representation to (a))	2.59	0.67	64.02	2.17	0.62	74.96
(f)	Ours (full, with all innovations in ablation studies, **best model)**	**2.51**	**0.62**	**73.41**	**2.07**	**0.47**	**79.32**

Table 7. Ablation study of our innovations on **office3** and **office4** of Replica [8].

	Methods	room0		room1		room2		office0		office1		office2		office3		office4	
		Acc.	mIoU	Acc.	mIoU	Acc.	mIoU	Acc.	mIoU	Acc.	mIoU	Acc.	mIoU	Acc.	mIoU	Acc.	mIoU
Offline	Semantic-NeRF [13]	98.34	97.00	98.71	97.28	97.59	95.92	98.67	97.44	97.07	93.48	97.62	92.42	96.74	94.31	96.62	94.22
Online	NIDS-SLAM [3]	97.76	82.45	98.50	84.08	98.76	76.99	98.89	85.94	–	–	–	–	–	–	–	–
	SNI-SLAM (Ours)	**98.53**	**88.42**	**98.61**	**87.43**	**98.80**	**86.16**	**99.17**	**87.63**	**99.18**	**78.63**	**99.13**	**86.49**	**98.66**	**74.01**	**99.06**	**80.22**

Table 8. Quantitative comparison of SNI-SLAM with existing semantic NeRF-based SLAM method NIDS-SLAM [3] and offline 3D semantic reconstruction method Semantic-NeRF [13]. For a fair comparison, the results are obtained using ground truth semantic labels for supervision. Online methods only take RGB-D frames as input, while offline method requires RGB-D frames and corresponding camera poses. Online methods perform scene reconstruction in several minutes while offline method require hours of training. Our method surpasses the performance of NIDS-SLAM [3] and achieves comparable results with Semantic-NeRF [13] in pixel accuracy metric.

	Methods	Reconstruction [cm]				Localization [cm]	
		Depth L1 ↓	Acc. ↓	Comp. ↓	Comp.Ratio(%) ↑	RMSE ↓	Mean ↓
office0	iMAP* [10]	3.79	3.34	3.62	83.59	3.31	2.74
	NICE-SLAM [14]	1.43	1.83	1.84	94.93	1.50	1.32
	Vox-Fusion [12]	3.44	1.63	1.87	93.86	1.35	0.98
	Co-SLAM [11]	1.24	1.57	1.56	96.09	0.69	0.63
	ESLAM [4]	0.71	1.61	1.45	98.45	0.61	0.45
	SNI-SLAM (Ours)	**0.55**	**1.46**	**1.30**	**98.70**	**0.33**	**0.28**
office1	iMAP* [10]	3.76	2.10	3.62	88.45	1.42	1.15
	NICE-SLAM [14]	1.58	1.76	1.82	94.11	1.01	0.91
	Vox-Fusion [12]	1.77	1.60	1.66	94.40	1.76	1.29
	Co-SLAM [11]	1.48	**1.31**	1.59	94.65	0.56	0.52
	ESLAM [4]	1.02	1.82	1.30	97.60	0.59	0.51
	SNI-SLAM (Ours)	**0.97**	1.58	**1.26**	**97.70**	**0.41**	**0.35**
office2	iMAP* [10]	3.97	4.06	4.73	79.73	7.17	4.81
	NICE-SLAM [14]	2.70	3.18	3.11	88.27	1.85	1.51
	Vox-Fusion [12]	3.52	**2.02**	3.03	88.94	1.18	0.73
	Co-SLAM [11]	1.86	2.84	2.43	91.63	2.12	1.98
	ESLAM [4]	0.93	2.95	1.92	95.07	0.67	0.50
	SNI-SLAM (Ours)	**0.89**	2.52	**1.87**	**95.20**	**0.32**	**0.28**
office3	iMAP* [10]	5.61	4.20	5.49	73.90	6.32	4.89
	NICE-SLAM [14]	2.10	3.01	3.16	87.68	5.67	2.53
	Vox-Fusion [12]	1.82	**2.33**	2.81	89.10	1.11	0.69
	Co-SLAM [11]	1.66	3.06	2.72	90.72	1.62	1.47
	ESLAM [4]	1.03	2.55	2.20	95.05	0.74	0.64
	SNI-SLAM (Ours)	**0.75**	2.51	**2.07**	**95.40**	**0.62**	**0.56**
office4	iMAP* [10]	5.71	4.34	6.65	74.77	2.55	2.10
	NICE-SLAM [14]	2.06	2.54	3.61	87.23	3.53	2.52
	Vox-Fusion [12]	4.84	**2.02**	3.51	86.53	1.64	1.18
	Co-SLAM [11]	1.54	2.23	2.52	90.44	0.87	0.68
	ESLAM [4]	1.18	2.10	2.13	94.31	0.66	0.54
	SNI-SLAM (Ours)	**0.97**	2.07	**2.10**	**94.40**	**0.47**	**0.40**
room0	iMAP* [10]	5.08	4.01	5.84	78.34	6.33	3.85
	NICE-SLAM [14]	1.79	2.44	2.60	91.81	1.86	1.49
	Vox-Fusion [12]	1.76	**1.77**	2.69	92.03	1.37	1.03
	Co-SLAM [11]	1.05	2.11	2.02	95.26	0.72	0.57
	ESLAM [4]	0.73	2.15	1.79	97.39	0.84	0.67
	SNI-SLAM (Ours)	**0.55**	2.09	**1.73**	**97.80**	**0.50**	**0.43**
room1	iMAP* [10]	3.44	3.04	4.40	85.85	3.46	2.91
	NICE-SLAM [14]	1.33	2.10	2.19	93.56	2.37	1.92
	Vox-Fusion [12]	2.52	**1.51**	2.31	92.47	1.90	1.35
	Co-SLAM [11]	0.85	1.68	1.81	95.19	0.85	0.73
	ESLAM [4]	0.74	1.94	1.58	96.50	0.72	0.58
	SNI-SLAM (Ours)	**0.58**	1.66	**1.56**	**96.74**	**0.55**	**0.48**
room2	iMAP* [10]	5.78	3.84	5.07	79.40	2.65	2.50
	NICE-SLAM [14]	2.20	2.17	2.73	91.48	2.26	1.65
	Vox-Fusion [12]	3.58	2.23	2.58	90.13	1.47	1.02
	Co-SLAM [11]	2.37	1.99	1.96	93.58	1.02	0.87
	ESLAM [4]	1.26	1.68	1.65	96.99	0.53	0.44
	SNI-SLAM (Ours)	**0.87**	**1.64**	**1.62**	**97.05**	**0.45**	**0.38**

Table 9. Per-scene mesh reconstruction and localization accuracy results in Replica dataset [8]. Best results are highlighted as **first**, second best results are highlighted as second. Our method achieves state-of-the-art performance in all scenes.

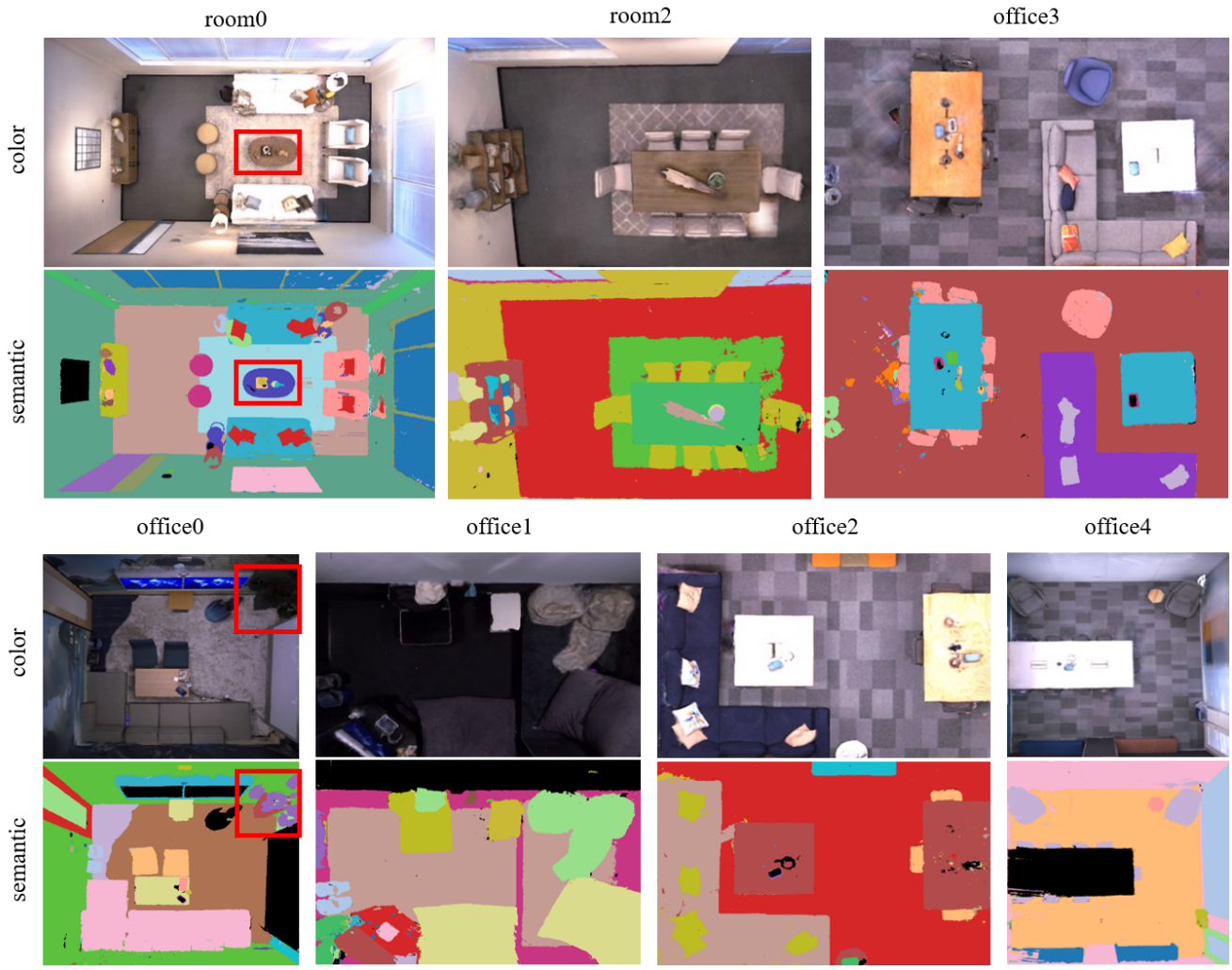

Figure 3. Top-view semantic reconstruction results of SNI-SLAM on the Replica dataset [8]. Our method is capable of achieving relatively accurate results through optimization during the real-time mapping process. As shown in red colored box of room0, small objects on the table are segmented accurately. Red colored box of office0 displays that flowers with complex shapes can be precisely segmented.

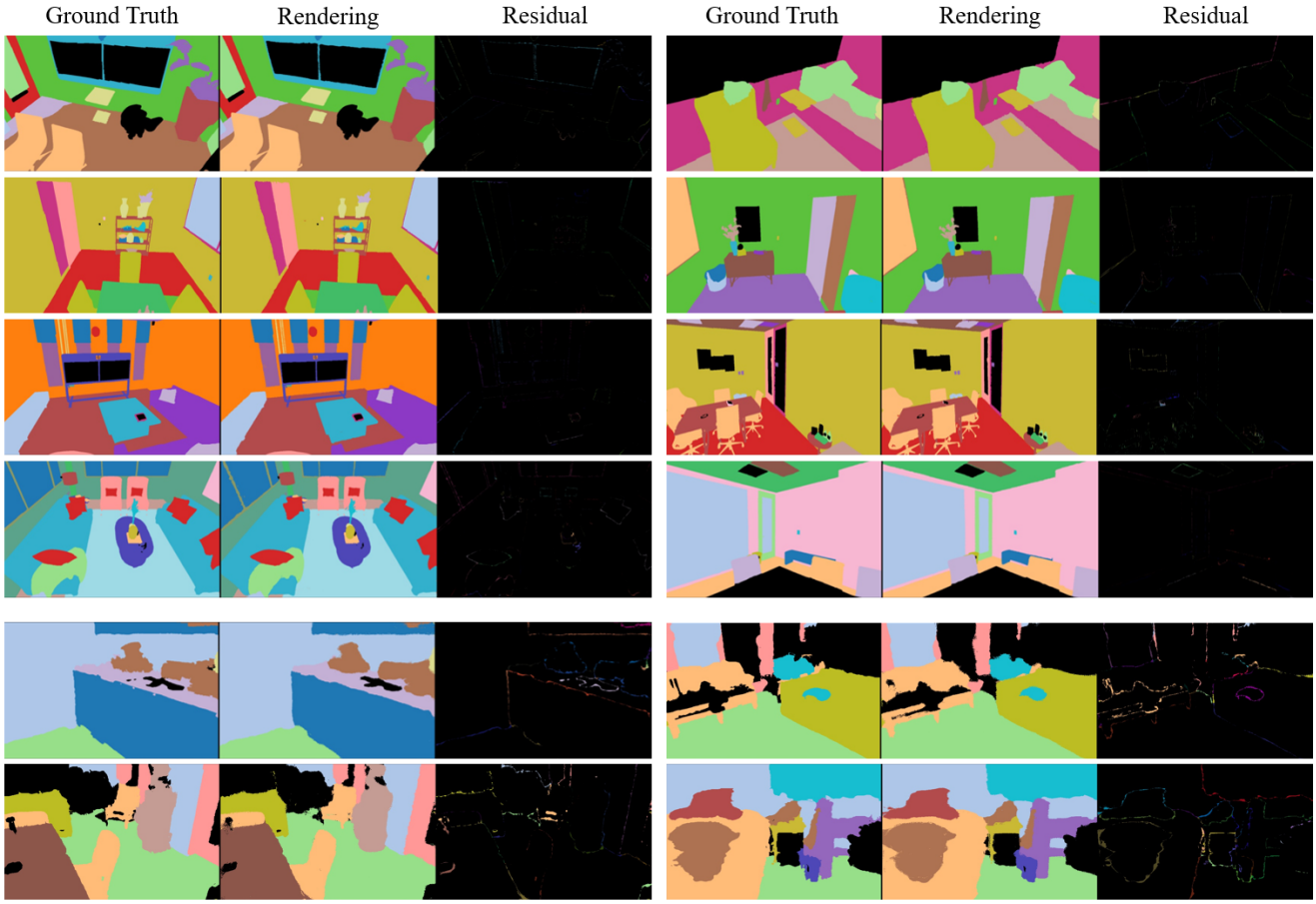

Figure 4. Semantic rendering results of SNI-SLAM on Replica [8] and ScanNet [2] datasets. Residual visualizes the difference between rendering results and the ground truth labels. It can be observed that our method can achieve excellent semantic segmentation accuracy.

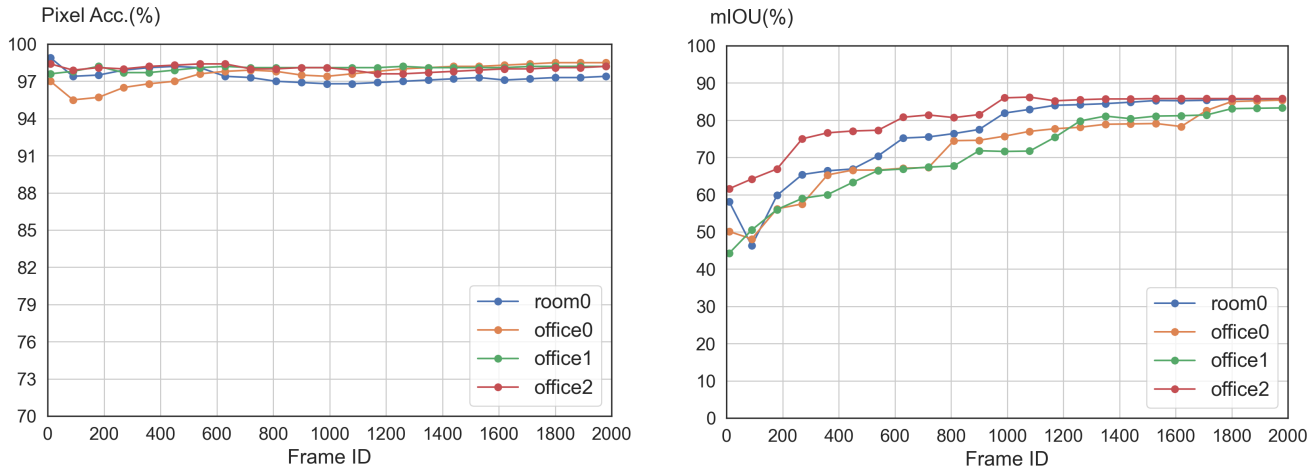

Figure 5. Qualitative results of SNI-SLAM in real-time semantic mapping from the first frame to the last frame. The y-axis represents pixel accuracy and mIOU of rendering labels, the x-axis represents frame index of simultaneously mapping and tracking. The graph displays that our method can already achieve high accuracy at the beginning of the mapping.

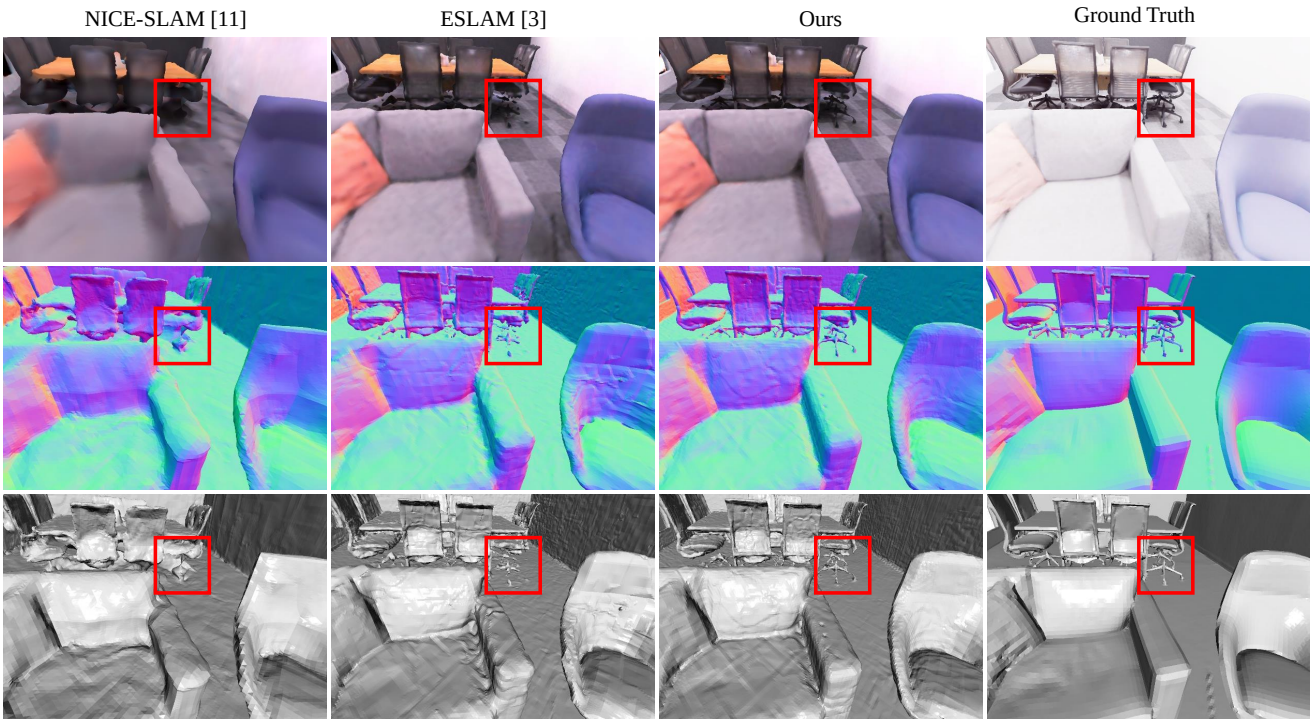

Figure 6. Qualitative comparison of our method with existing NeRF-based SLAM methods on Replica [8] of office3 using different shading mode. As shown in red colored box, other methods cannot accurately model chair legs while our method can. Moreover, our method achieves more accurate surface reconstruction results than baseline.

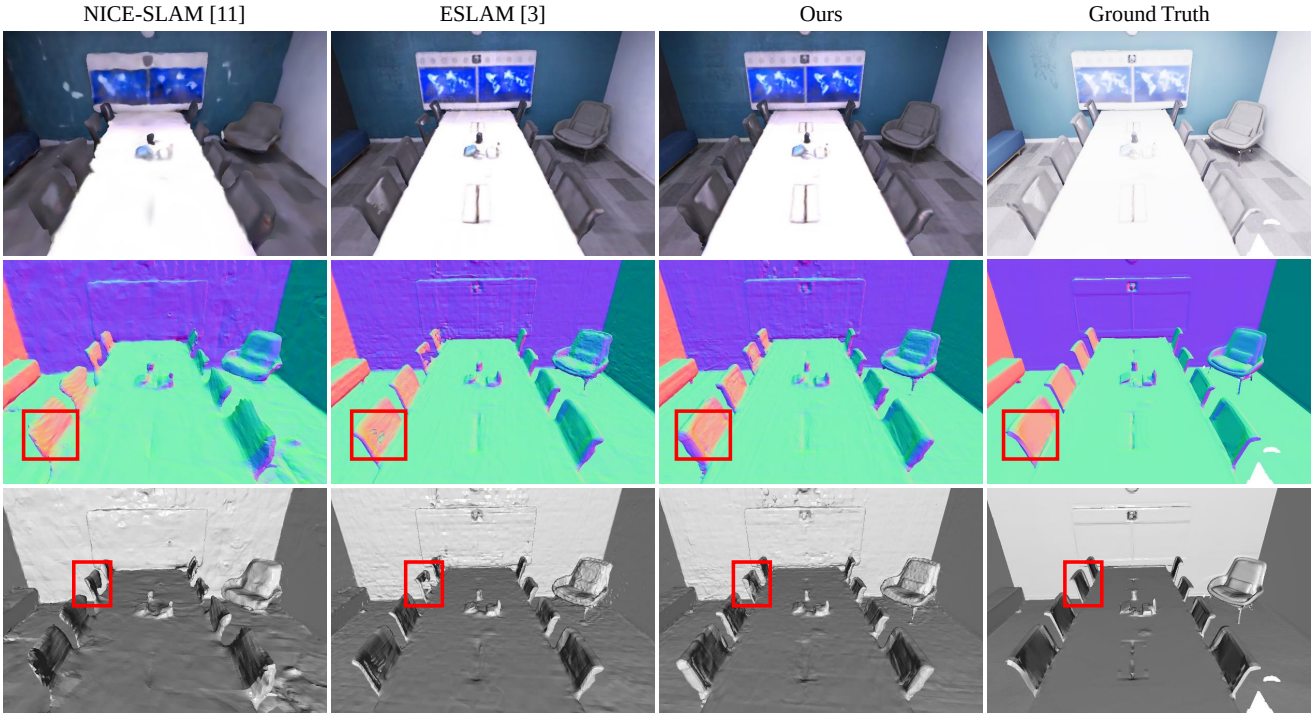

Figure 7. Qualitative comparison of our method with existing NeRF-based SLAM methods on Replica [8] of office4 using different shading mode. As shown in red colored box, our method achieves complete reconstruction compared with other methods.

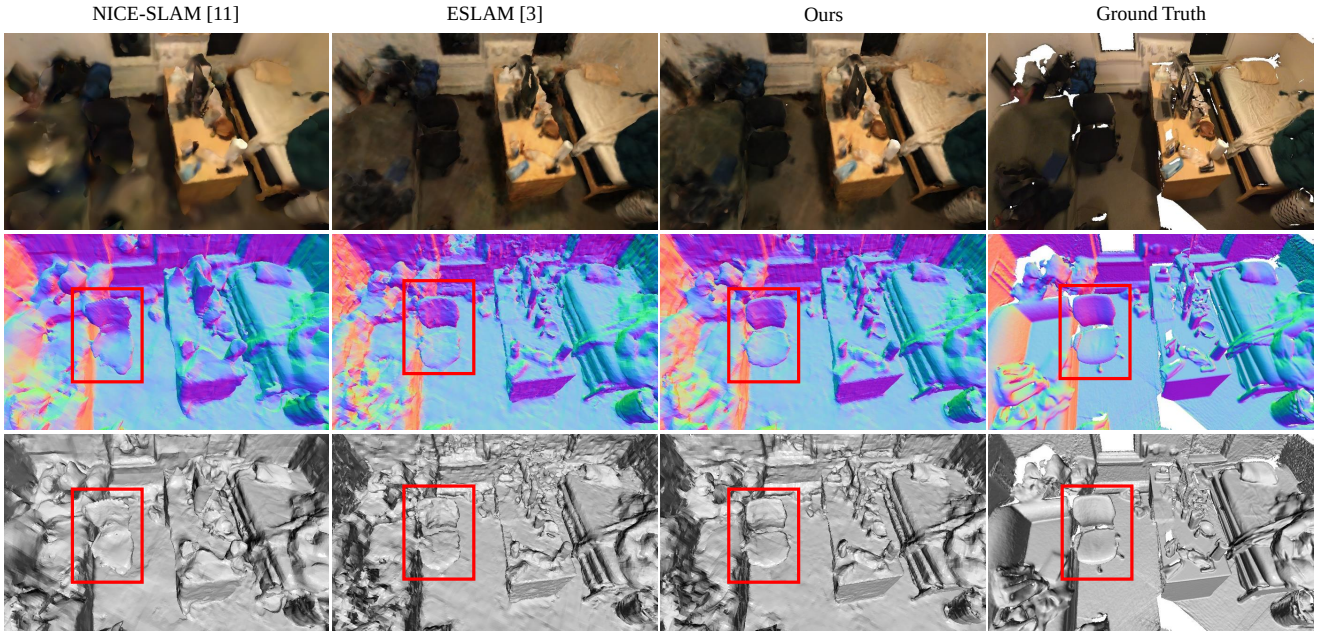

Figure 8. Qualitative comparison of our method with existing NeRF-based SLAM methods on ScanNet [2] of scene0207 using different shading mode. As shown in red colored box, our method achieves more accurate reconstruction compared with other methods.