Supplementary Material
Co-SLAM: Joint Coordinate and Sparse Parametric Encodings for Neural Real-Time SLAM

Hengyi Wang* Jingwen Wang* Lourdes Agapito
Department of Computer Science, University College London
{hengyi.wang.21, jingwen.wang.17, l.agapito}@ucl.ac.uk

1. Implementation details

1.1. Hyperparameters

Here we report the detailed settings and hyper-parameters used in Co-SLAM to achieve high-quality reconstruction and quasi-realtime performance.

Default setting. For camera tracking, we select $N_t = 1024$ pixels and perform 10 iterations of tracking with $M_c = 32$ regular sampling and $M_f = 11$ depth-guided sampling for each camera ray. In terms of mapping and bundle adjustments, we select $N_g = 2048$ pixels and use 200 iterations for first frame mapping, 10 iterations for bundle adjustment every 5 frames. For scene representations, we use $L = 16$ level HashGrid with from $R_{min} = 16$ to R_{max}, where we use max voxel size 2cm for determining R_{max}, and 16 bins for OneBlob encoding of each dimension. Two 2-layer shallow MLPs with 32 hidden units are used to decode color and SDF. The dimension of the geometric feature $\mathbf{h}$ is 15. For the training of our scene representation, we use learning rate of $1e{-}3$ for tracking and $1e{-}2$, $1e{-}2$, $1e{-}3$ for feature grid, decoder, and camera parameters during bundle adjustment. The weights of each loss are $\lambda_{rgb} = 5$, $\lambda_d = 0.1$, $\lambda_{sdf} = 1000$, $\lambda_{fs} = 10$, and $\lambda_{smooth} = 1e{-}6$. The truncation distance tr is set to 10cm.

ScanNet dataset. For ScanNet dataset, we change the voxel size to $4cm$ for the finest resolution, and increase the number of sample points to $M_c = 96$, $M_f = 21$. The λ_{smooth} is increased to $1e-3$.

TUM dataset. Since scenes in TUM dataset is mostly focusing on reconstructions of tables instead of the whole room, we use 20 iterations for bundle adjustment, and set tr to 5cm. The weights of each loss are $\lambda_{rgb} = 1.0$, $\lambda_d = 0.1$, $\lambda_{sdf} = 5000$, $\lambda_{fs} = 10$, and $\lambda_{smooth} = 1e{-}8$. The learning rate of camera parameters in tracking process is increased to $1e-2$.

* Indicates equal contribution.

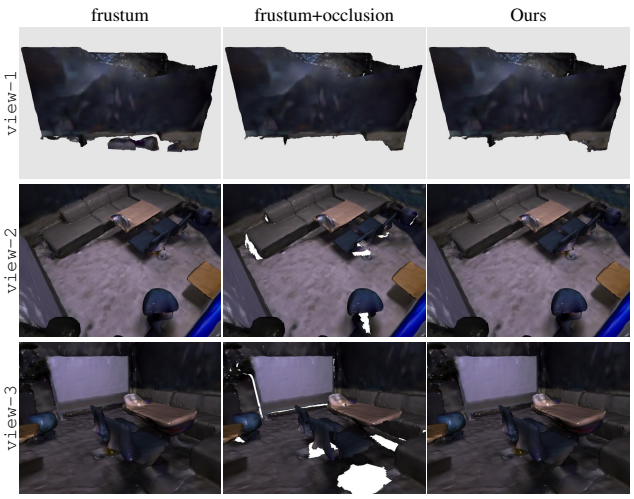


Figure 1. Visualization of different culling strategy applied on mesh reconstructed by NICE-SLAM [7]. Culling by frustum fails to remove the artefacts outside the scene bound (`view-1`), while culling by frustum+occlusion removes occluded regions (`view-2` and `view-3`) inside the room. Our proposed method could remove unwanted artefacts outside the room but preserve the completeness inside the room.

1.2. Evaluation Protocol

We have introduced a modification to the culling strategy used for the quantitative evaluation of the reconstruction accuracy, which we believe leads to a fairer comparison. In this section we describe this new culling strategy in detail and provide a justification for its use.

In the context of neural implicit reconstruction and SLAM [1, 5–7], due to the extrapolation ability of neural networks an extra mesh culling step is required before evaluating the reconstructed mesh. We show a demonstration in Fig. 1. In previous works two different culling strategies are used: NICE-SLAM [7] and iMAP [5] adopted a *culling-by-frustum* strategy where mesh vertices outside any of the camera frustums are removed. This simple strategy works effectively well but cannot remove artifacts that are

	iMAP* [5]	NICE-SLAM [7]	Ours
Depth L1 ↓	7.64	3.53	**1.58**
Acc. ↓	6.95	2.85	**2.15**
Comp. ↓	5.33	3.00	**2.21**
Comp. Ratio ↑	66.60	89.33	**92.99**

Table 1. Reconstruction results on Replica dataset using NICE-SLAM culling strategy. The smooth weight λ_{smooth} is increased to $1e-3$, and hash table size is set to be 14.

3D Metric	definition
Acc	$\sum_{p\in P}(\min_{q\in Q}\|p-q\|)/\|P\|$
Comp	$\sum_{q\in Q}(\min_{p\in P}\|p-q\|)/\|Q\|$
C-ℓ_1	$(\text{Acc}+\text{Comp})/2$
Comp Ratio	$\sum_{q\in Q}(\min_{p\in P}\|p-q\|<0.05)/\|Q\|$

Table 2. Definitions of 3D metrics used for evaluation of reconstruction quality.

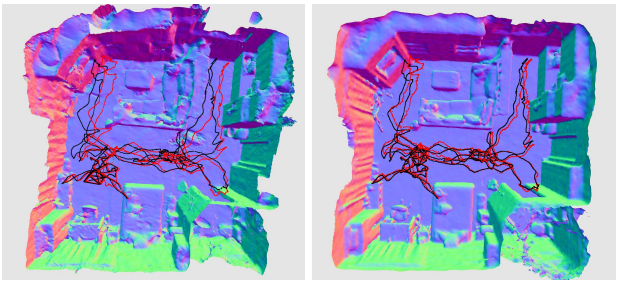
inside camera frustums but outside of scene bound, such as in `view-1`. In NeuralRGBD [1] and GO-Surf [6] the *frustum+occlusion* strategy is used, where in addition to the frustum criteria self-occlusion is also considered by comparing the rendered depth. While this strategy could effectively remove the artifacts in `view-1` it also removes the occluded regions as in `view-2` and `view-3`. Therefore, we propose a new culling strategy that follows the *frustum+occlusion* criteria but also simulates virtual camera views that cover the occluded regions. Since we focus on the inner surface of the scene in Replica dataset [4], we can remove the noisy points outside the mesh of interest to make fair comparison of different methods.

We also show evaluation of reconstruction quality using nice-slam culling strategy in Tab. 1. To remove the noisy points outside the outer surface caused by hash collision, we increase the smooth weight to $\lambda_{smooth} = 1e-3$ and increase the default hash lookup table size from 13 to 14.

1.3. Evaluation Metrics

After mesh culling we evaluate the reconstructed mesh with a mixture of 3D (**Acc**uracy, **Comp**letion and **Comp**letion **Ratio**) and 2D (**Depth L1**) metrics. We first uniformly sample two point clouds P and Q from both GT and reconstructed meshes, with $|P| = |Q| = 200000$. Then accuracy is defined as the average distance between a point on GT mesh to its nearest point on reconstructed mesh, other metrics are defined in the same fashion, see Tab. 2.

For depth L1, following [7] we render depth from $N = 1000$ virtual views of GT and reconstructed mesh. The virtual views are sampled uniformly inside a cube within the room. Views that have unobserved points will be rejected and re-sampled. Then depth L1 is defined as the average L1 difference between rendered GT depth and reconstruction depth.



NICE-SLAM Ours

Figure 2. A more comprehensive comparison of raw trajectory w/o alignment. In both figures ground truth trajectory are shown in black and the estimated trajectory are shown in red.

Scene ID	0000	0059	0106	0169	0181	0207	Avg.
NICE-SLAM	8.64	12.25	**8.09**	10.28	12.93	**5.59**	9.63
Ours	**7.13**	**11.14**	9.36	**5.90**	**11.81**	7.14	**8.75**
NICE-SLAM	25.24	25.01	**10.40**	30.51	39.98	12.70	23.97
Ours	**12.94**	**19.12**	12.12	**19.61**	**34.41**	**9.88**	**18.01**

Table 3. ATE RMSE (cm) results (w/ and w/o trajectory alignment) of average of 5 runs on ScanNet.

2. Additional Experimental Results

2.1. More Results on Synthetic Datasets

Here we show more detailed results on all the synthetic scenes in Replica [4] and Synthetic RGB-D [1]. We show per-scene quantitative results of the Replica [4] and Synthetic RGB-D dataset [1] in Tab. 4 and Tab. 5. Our method shows consistently better results in terms of the *Completion* and competitive results of *Accuracy*. We also provide more qualitative comparisons on Replica dataset in Fig. 4-7 in different scenes with different shading mode.

2.2. More Results on Real-world Scenes

ScanNet sequences [2]. In our main paper we showed top-down view comaparison on ScanNet sequence, which highlighted more on overall reconstruction quality and tracking accuracy. In this section we show more detailed zoom-in views in Fig. 8-10 to better showcase the level of details and fidality that Co-SLAM can achieve on those challenging real-world sequences.

NICE-SLAM apartment [7]. In addition to the 6 sequences from ScanNet, we also compared Co-SLAM and NICE-SLAM on the apartment sequence collected by the authors of NICE-SLAM using Azure Kinect depth camera. We run Co-SLAM with our ScanNet setting. We show qualitative comparison from different views in Fig. 11. As can be seen Co-SLAM achieves smoother and better quality reconstruction in much shorter time (40 minutes vs. 10 hours).

Self-captured room. In addition to ScanNet sequences and the apartment sequences captured by the authors of NICE-SLAM, we also collected our two real-world indoor se-

		room0	room1	room2	office0	office1	office2	office3	office4	Avg.
iMAP	**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 [< 5cm %] ↑	78.34	85.85	79.40	83.59	88.45	79.73	73.90	74.77	80.50
NICE-SLAM	**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 [< 5cm %] ↑	91.81	93.56	91.48	94.93	94.11	88.27	87.68	87.23	91.13
Co-SLAM	**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 [< 5cm %] ↑	**95.26**	**95.19**	**93.58**	**96.09**	**94.65**	**91.63**	**90.72**	**90.44**	**93.44**

Table 4. Per-scene quantitative results on Replica [4] dataset. Our method achieves consistently better reconstruction in comparison to NICE-SLAM [7] and iMAP [5] in most of the scenes.

		BR	CK	GR	GWR	MA	TG	WR	Avg.
iMAP*	**Depth L1** [cm] ↓	24.03	63.59	26.22	21.32	61.29	29.16	81.71	47.22
	Acc. [cm] ↓	10.56	25.16	13.01	11.90	29.62	12.98	24.82	18.29
	Comp. [cm] ↓	11.27	31.09	19.17	20.39	49.22	21.07	32.63	26.41
	Comp. Ratio [< 5cm %] ↑	46.91	12.96	21.78	20.48	10.72	19.17	13.07	20.73
NICE-SLAM	**Depth L1** [cm] ↓	3.66	12.08	10.88	2.57	1.72	7.74	5.59	6.32
	Acc. [cm] ↓	3.44	10.92	5.34	2.63	6.55	3.57	9.22	5.95
	Comp. [cm] ↓	3.69	12.00	4.94	3.15	3.13	5.28	4.89	5.30
	Comp. Ratio [< 5cm %] ↑	87.69	55.41	82.78	87.72	85.04	72.05	71.56	77.46
Co-SLAM	**Depth L1** [cm] ↓	**3.51**	**5.62**	**1.95**	**1.25**	**1.41**	**4.66**	**2.74**	**3.02**
	Acc. [cm] ↓	**1.97**	**4.68**	**2.10**	**1.89**	**1.60**	**3.38**	**5.03**	**2.95**
	Comp. [cm] ↓	**1.93**	**4.94**	**2.96**	**2.16**	**2.67**	**2.74**	**3.34**	**2.96**
	Comp. Ratio [< 5cm %] ↑	**94.75**	**68.91**	**90.80**	**95.04**	**86.98**	**86.74**	**84.94**	**86.88**

Table 5. Per-scene quantitative results on Synthetic RGBD [1] dataset. Since this dataset simulates noisy depth maps with missing depth measurement, our method surpasses NICE-SLAM [7] by a larger margin. This indicates our method is more robust to input noise.

quences using RealSense D435i depth camera, whose depth quality is slightly worse than Azure Kinect. We show qualitative comparison in Fig. 12 and Fig. 13.

2.3. ScanNet Camera Tracking Results

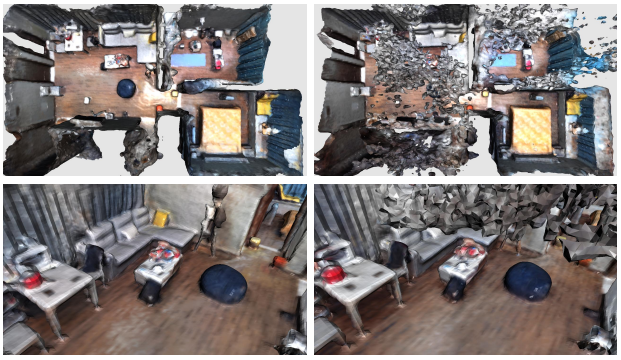
In this section, we provide a more comprehensive view of the camera tracking results on the ScanNet dataset. In evaluating the absolute trajectory error (ATE), a rigid transformation is estimated to align the estimated trajectory with the ground truth. While this protocol is widely used in traditional SLAM and also in NICE-SLAM [7], we observe that this does not always tell the whole story. For example, Fig. 2 shows the reconstructed ScanNet scene with estimated camera trajectory under the same world coordinate, i.e. **without** doing the trajectory alignment. It can be seen that Co-SLAM performs better in terms of camera tracking (Note the top-left and top-right corner of the trajectory) and leads to less distorted reconstruction. However, this is not reflected in Tab. 4 in our main paper as trajectory alignment. Therefore in Tab. 3 we report the full camera tracking results both with and without trajectory alignment. As can be seen, Co-SLAM achieves overall better and more robust tracking results.

2.4. More Ablation Studies

Using separate color grid. As dedcribed om our main paper, we adopted two separate MLP decoders for color and geometry but only use a single hash-grid. In Tab 6 we shows the comparison of using two separate hash grids for color and geometry, and using one hash grid (our default setting). We empirically discover that thanks to our joint coordinate and sparse parametric encoding, using a single hash-grid already achieves similar tracking accuracy and reconstruction quality while is more computational efficient and requires much less memory storage.

Effect of smoothness term. We also conduct ablation study on the effectiveness of our smoothness term applied on the features. As shown in Fig. 3, our smooth loss could provide a effective regularisation and remove artefacts caused by hash collisions in unobserved regions that do not have any supervision.

Effect of pose optimization in global BA. We perform an additional experiment on performing our GBA without any pose optimization (GBA‡) in Tab. 8. We show that even the sample points in each sampled batch (2048 rays) may not have large overlapping, optimizing camera pose with



w/ smooth term w/o smooth term

Figure 3. Reconstruction result w/ and w/o our smoothness term. Smoothness term could effectively remove hash collision artefacts.

Method	Tracking (ms)	Mapping (ms)	Memory	ATE↓
Two grids	9.9×20	25.4×10	1.7M	8.69
One grid	7.8×20	20.2×10	0.8M	8.75

Table 6. Performance analysis of modeling the geometry and appearance using one/two grids. The ATE is similar while using one grid require less computational cost. Run-time is reported in `ms/iter × #iter` format.

Method	**Acc.** ↓	**Comp.** ↓	**C-ℓ_1** ↓	**NC** ↑	**F-score** ↑	**Run time**
Neural RGBD	0.0151	0.0197	0.0174	0.9316	**0.9635**	10-25h
GO-Surf	0.0158	0.0195	0.0177	**0.9317**	0.9591	15-45min
Ours	**0.0149**	**0.0179**	**0.0164**	0.9292	0.9629	100-500s

Table 7. Quantitative results of the reconstruction on 10 synthetic scenes [1]. The evaluation metrics and protocol follow Neural RGBD [1] and GO-Surf [6]. We achieve on-par performance but our training is significantly faster.

Name	KF selection		#KF			Pose optim.	ATE (cm)	Std. (cm)
	Local	Global	0	10	All			
w/o BA			✓				16.81	1.69
LBA	✓			✓		✓	9.69	1.38
GBA-10		✓		✓		✓	9.54	0.67
GBA‡		✓			✓		9.72	0.53
GBA		✓			✓	✓	**8.75**	**0.33**

Table 8. Ablation using different BA strategies on Co-SLAM: (LBA) BA with rays from 10 local keyframes; (GBA-10) BA with rays from 10 randomly selected keyframes; (GBA‡): BA with rays from all keyframes w/o pose optimization; (GBA) BA with rays from all keyframes and pose optimization (our full method). All methods sample a total of 2048 rays per iteration.

such sampling strategy can still significantly improve the performance and the robustness of the pose estimation.

2.5. Batch-mode Optimisation

To validate the representation ability of the proposed joint coordinate and parametric encoding, we also perform experiments of batch mode optimisation, which is an offline approach and the pose initialised by BundleFusion [3] is given. Tab. 7 shows the quantitative results of different methods. Neural RGBD [1] is a coordinate encoding-based method while GO-Surf [6] is a parametric encoding-based method. By using the proposed joint coordinate and parametric encoding, we achieve competitive reconstruction performance with significantly faster training speed.

References

[1] Dejan Azinović, Ricardo Martin-Brualla, Dan B. Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6290–6301, 2022. 1, 2, 3, 4

[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. 2

[3] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017. 4

[4] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, and Shobhit Verma. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 3

[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. 1, 2, 3, 5, 6, 7, 8

[6] Jingwen Wang, Tymoteusz Bleja, and Lourdes Agapito. Go-surf: Neural feature grid optimization for fast, high-fidelity rgb-d surface reconstruction. *arXiv preprint arXiv:2206.14735*, 2022. 1, 2, 4

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 1, 2, 3, 5, 6, 7, 8

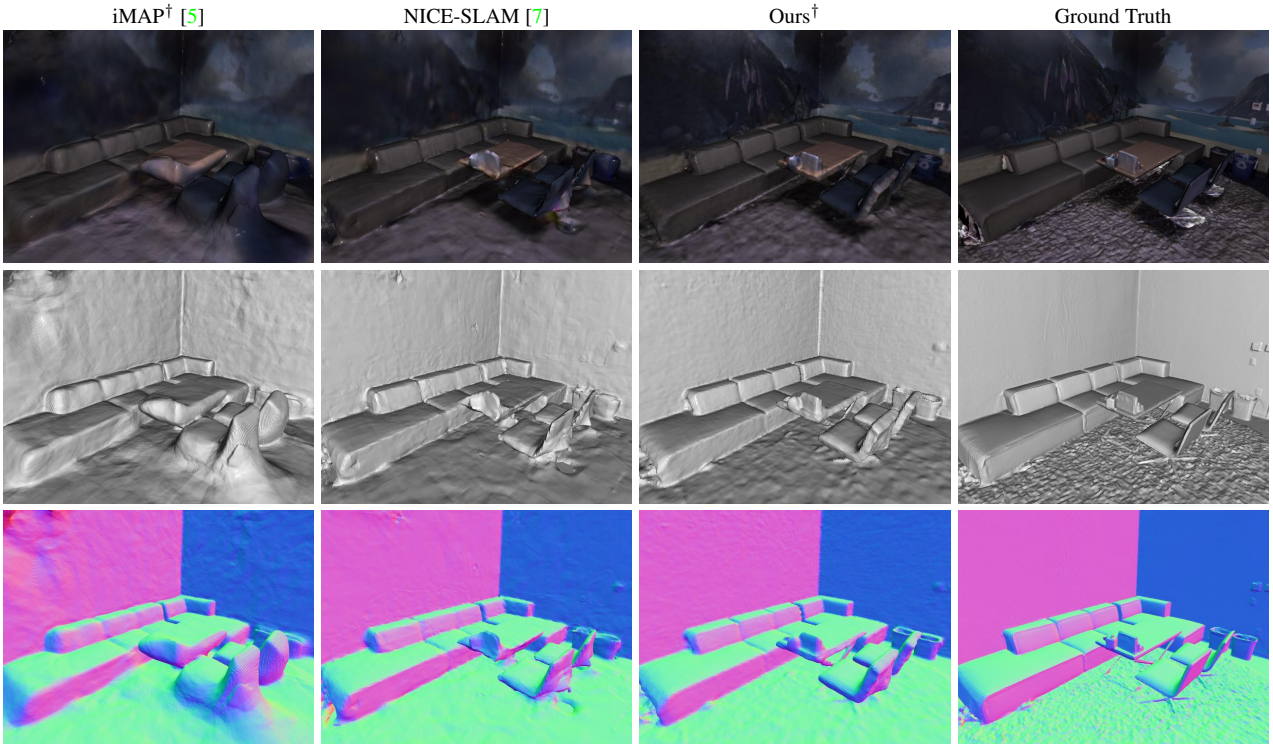


Figure 4. Qualitative comparison on Replica `office-0` with different shading mode. Our methods achieve accurate scene reconstruction with high frequence details. At the same time, our reconstruction is also sharper and smoother.

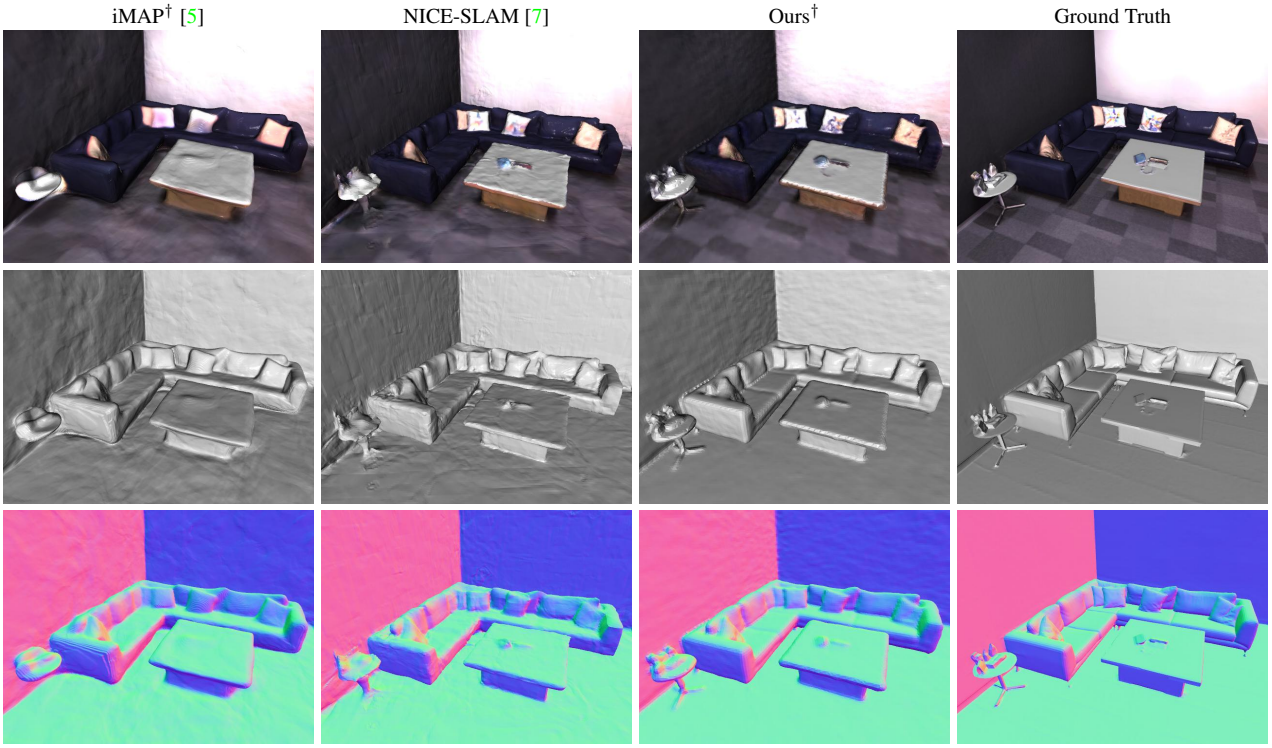


Figure 5. Qualitative comparison on Replica `office-2` with different shading mode. Note that regions with different color styles in the groundtruth color image indicate the unobserved region. Our method achieve better scene completion for unobserved regions.

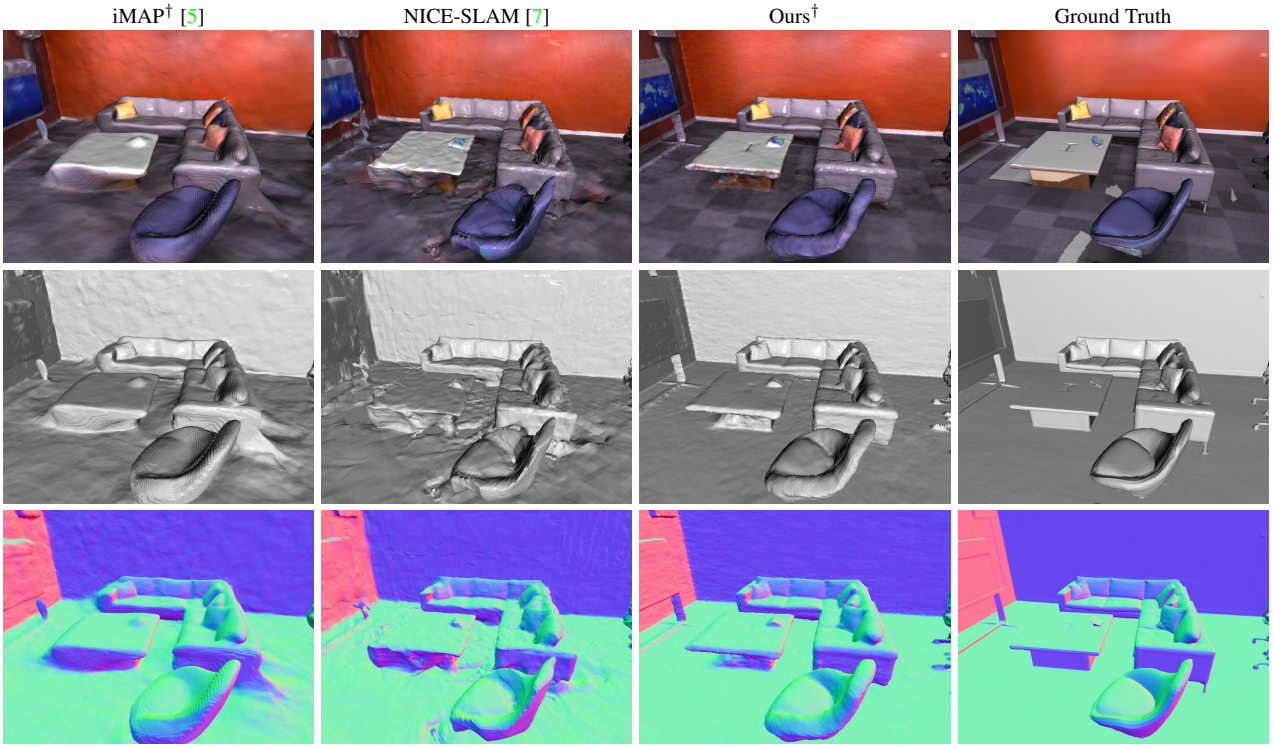


Figure 6. Qualitative comparison on Replica `office-3` with different shading mode. Our method achieves smooth reconstruction for regions that contain multiple objects while other methods contain some build-up effect.

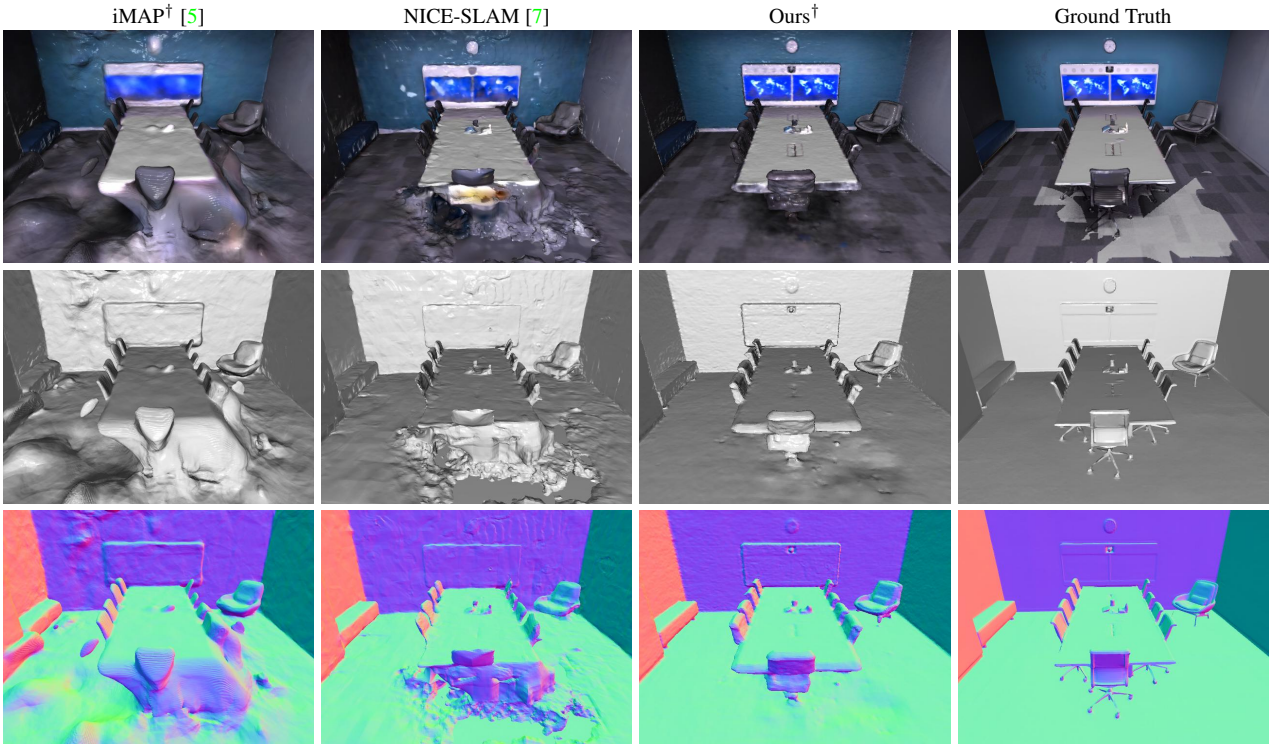


Figure 7. Qualitative comparison on Replica `office-4` with different shading mode. Note that regions with different color styles in the groundtruth color image indicate the unobserved region. Our method can accurately recover the thin structures while achieve smooth reconstruction around the flat regions that have not been observed.

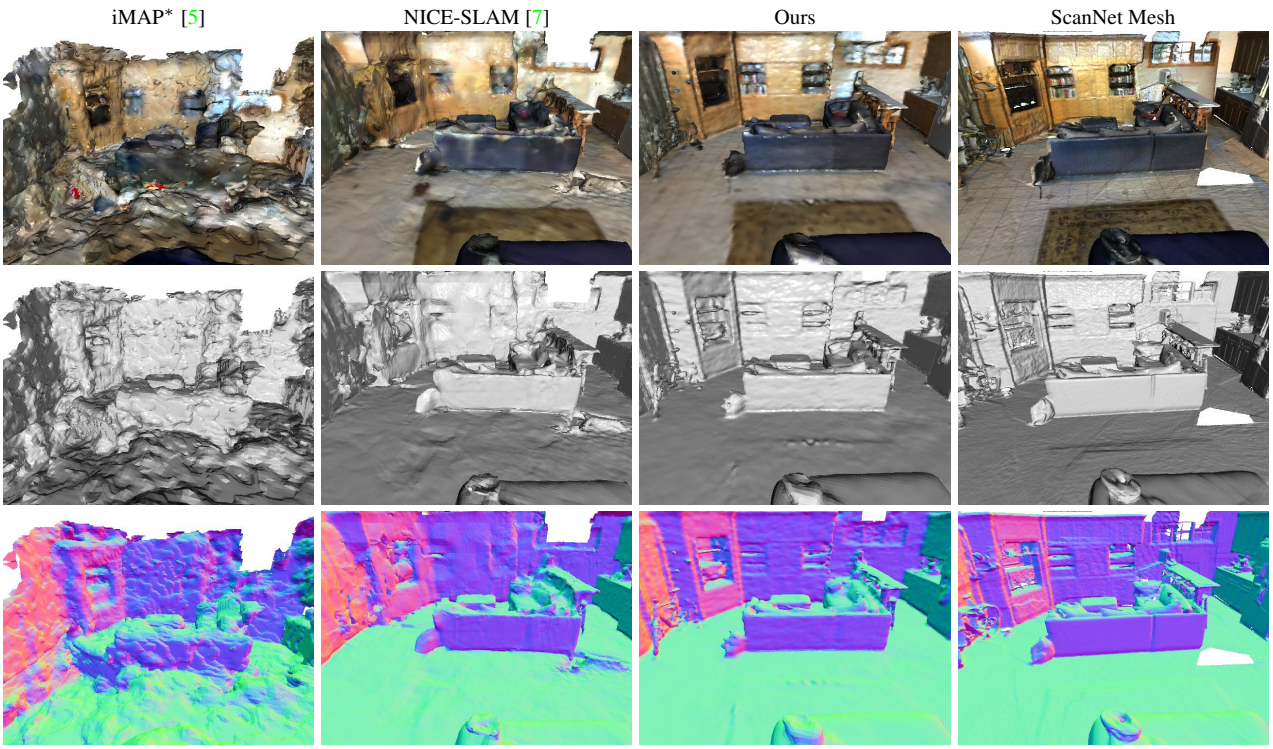


Figure 8. Qualitative comparison on ScanNet `scene0000` with different shading mode. For real world scans, our method can accurately recover thin structures (e.g. bicycle) while achieve better hole fillings. Since we adopt global bundle adjustment, our scene reconstruction seems to have better coherence while reconstruction of NICE-SLAM [7] contains some stitched effect due to the local bundle adjustment.

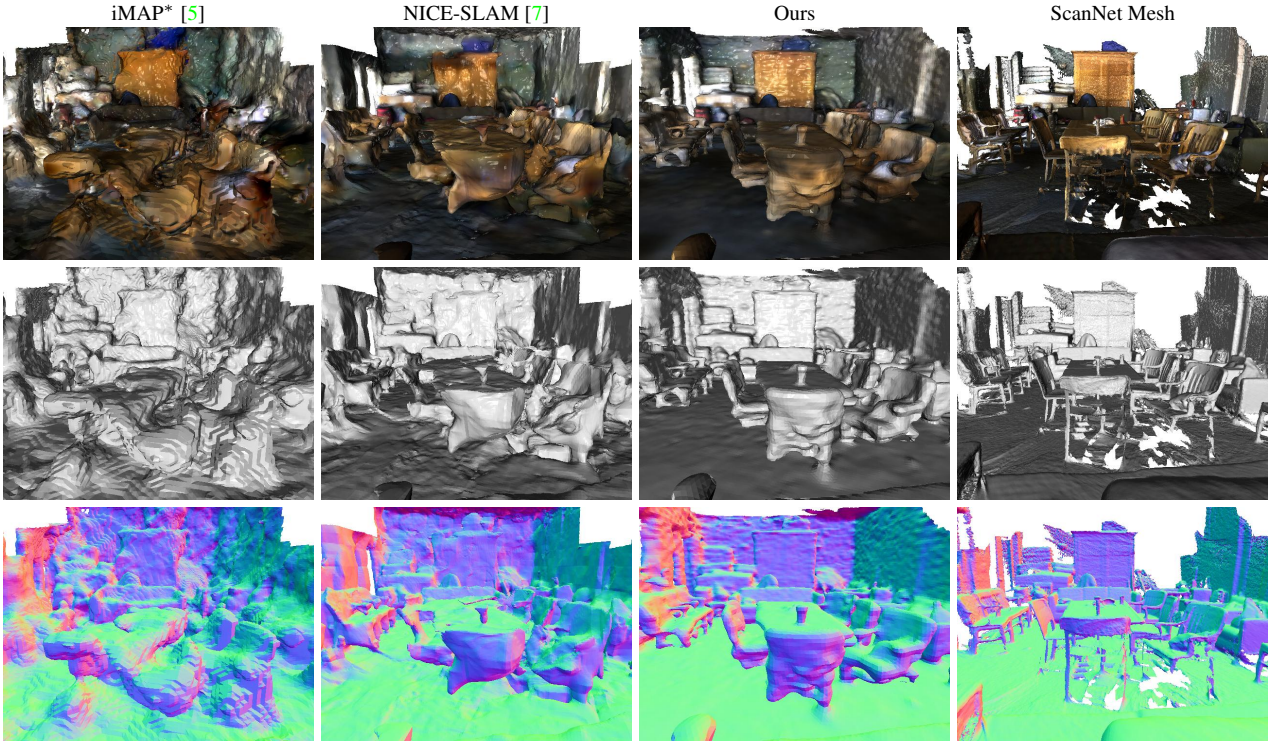


Figure 9. Qualitative comparison on ScanNet `scene0059` with different shading mode. Our method achieve smooth reconstruction of the floor while accurately recover the chairs.

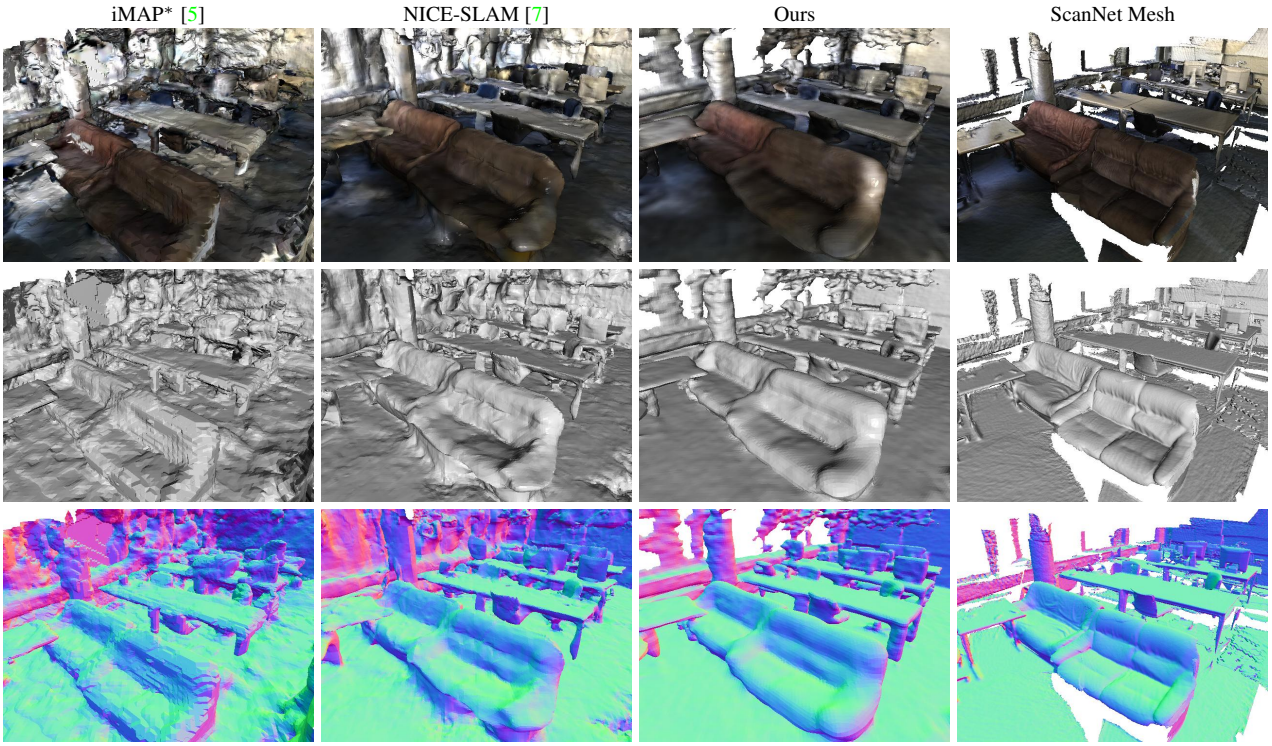


Figure 10. Qualitative comparison on ScanNet `scene0106` with different shading mode. Our reconstruction result is clearly less noisy in comparison to other two baseline models.

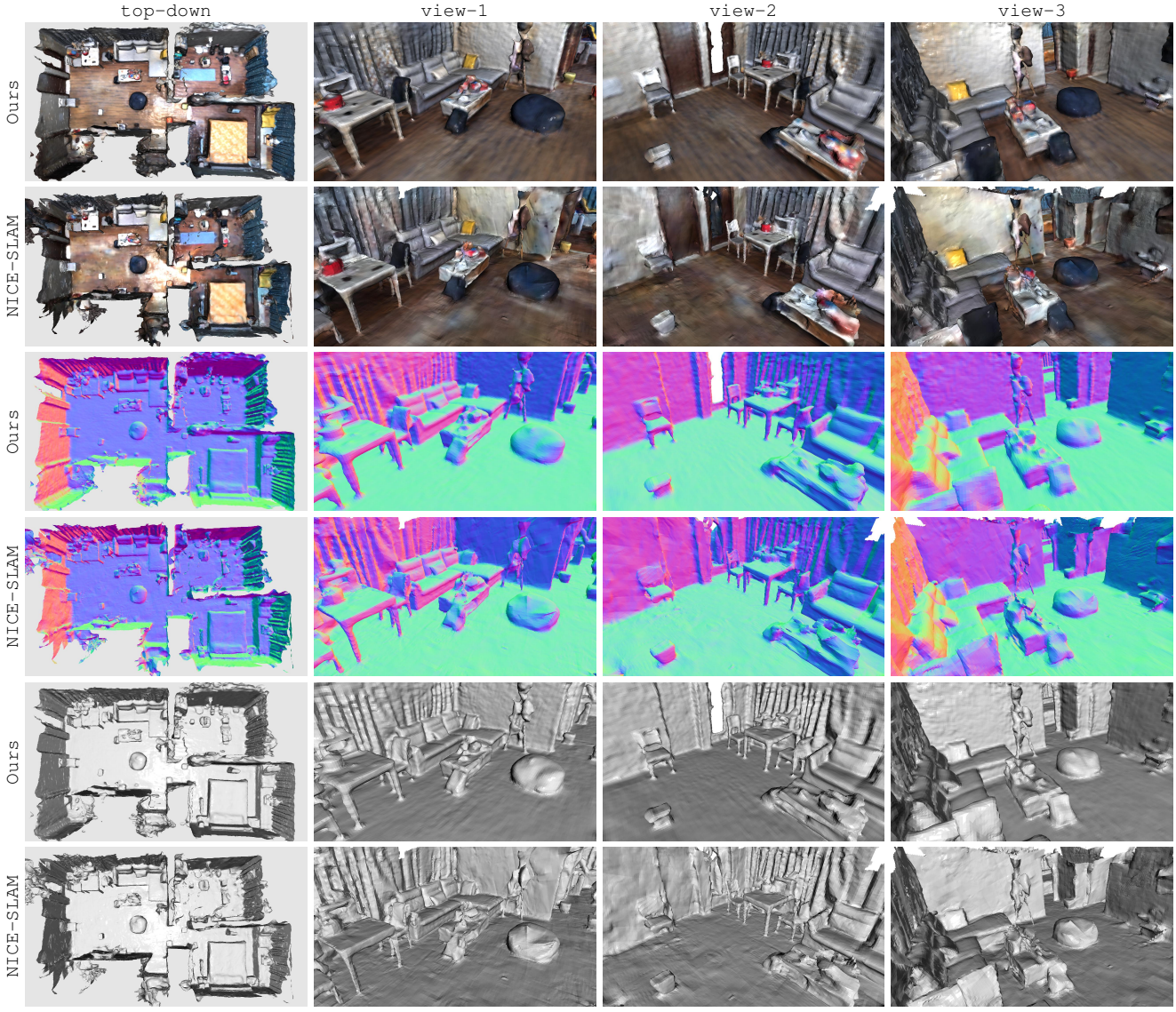


Figure 11. Qualitative comparison on NICE-SLAM apartment sequence on different view-point with different shading mode. Co-SLAM achieves smooth, detailed and high-fidelity reconstruction while running $>$ 10 times faster.

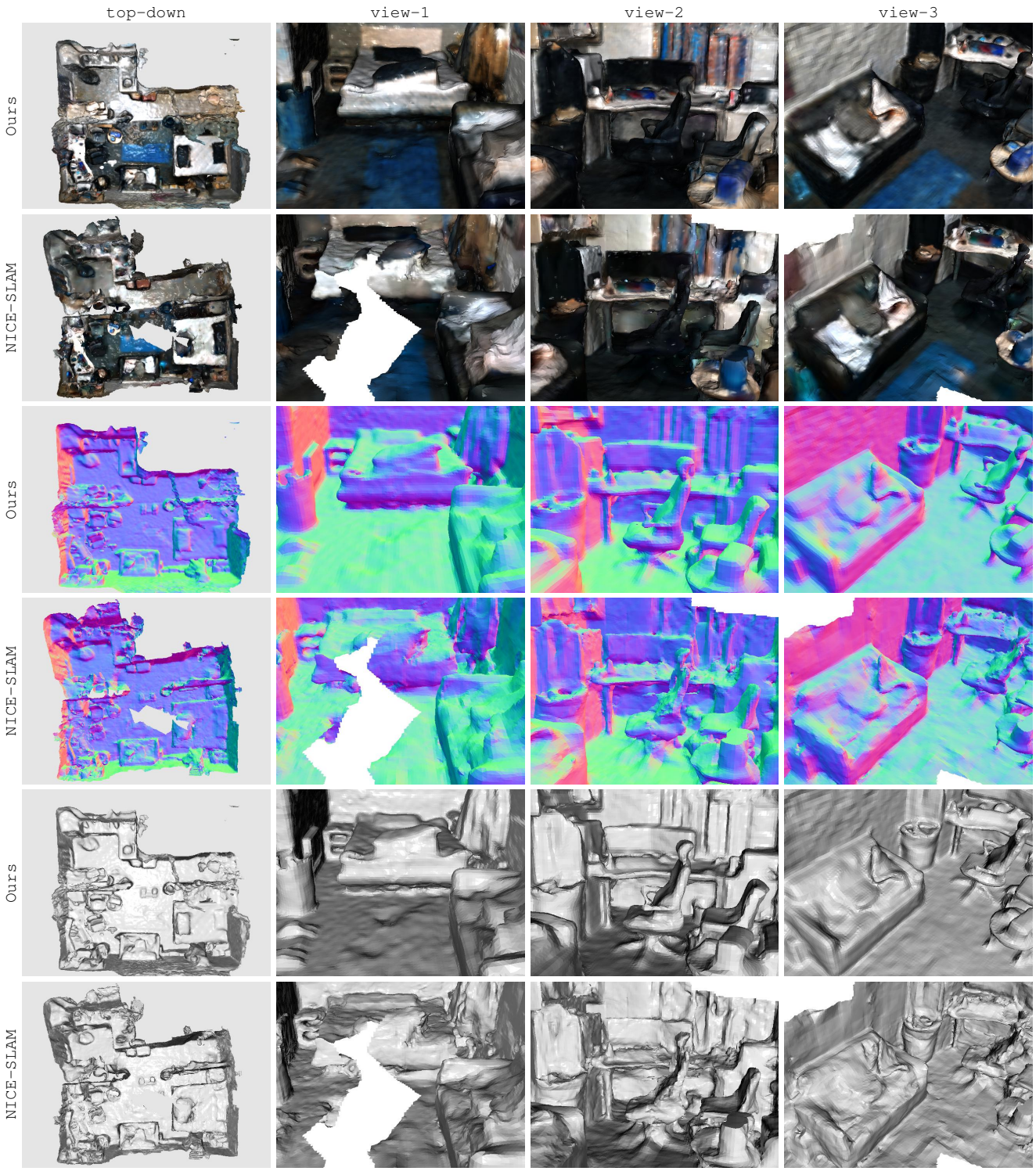


Figure 12. Qualitative comparison on self-captured room sequence on different view-point with different shading mode. Overall Co-SLAM produces higher quality surface reconstruction with finer details (the desk chair, the objects on the desk and sofa, the curtain, etc). Also note that NICE-SLAM lost tracking slightly causing the reconstructed scene to be torn apart.

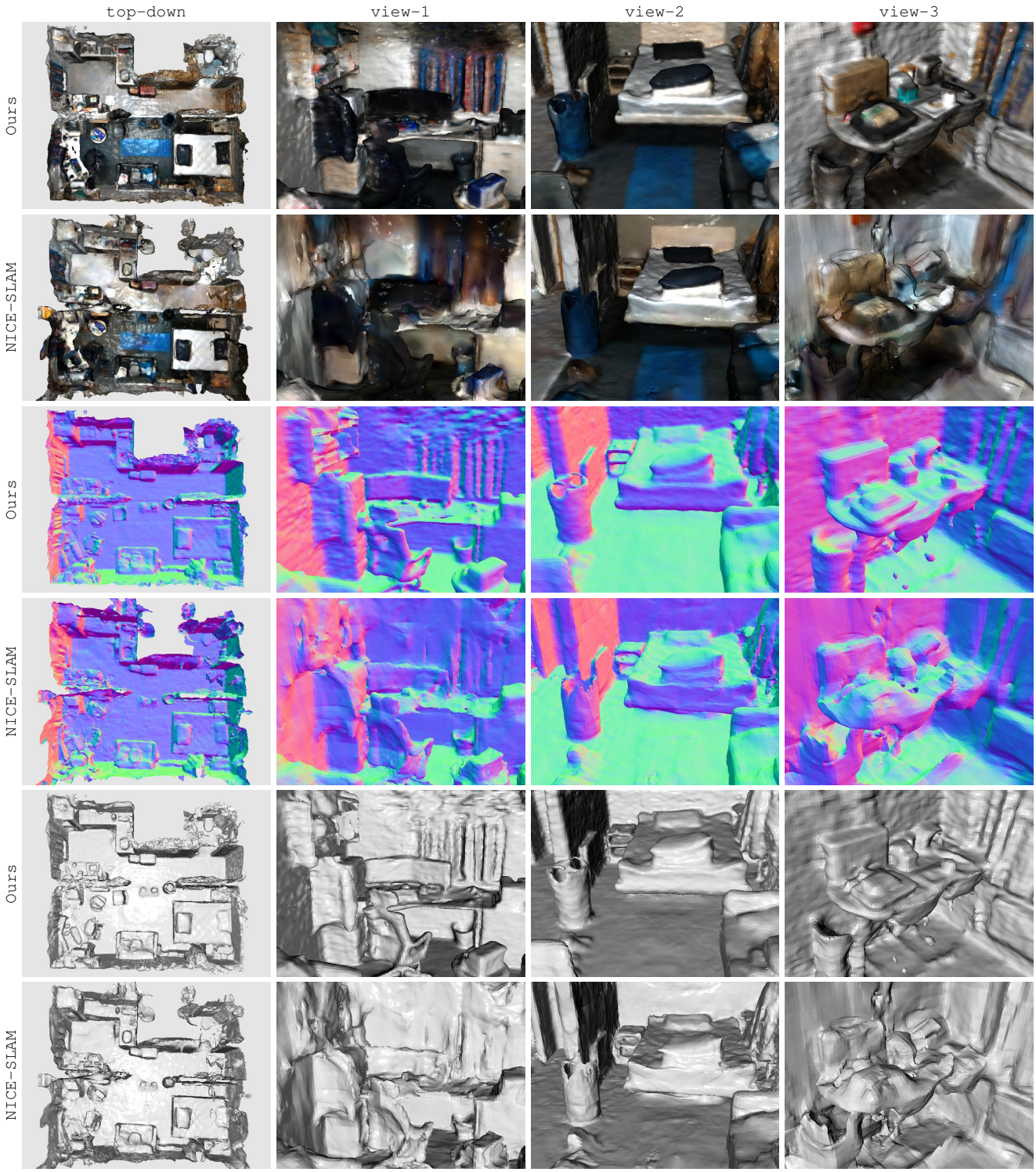


Figure 13. Qualitative comparison on a different scan of the same room in Fig. 12. Note how Co-SLAM produces better surface reconstruction while running $>$ 10 time faster.