# IBD-SLAM: Learning Image-Based Depth Fusion for Generalizable SLAM

## Supplementary Material

## A. Implementation Details

**Reference Frame Selection.** To render a new frame, we need to select several frames with estimated camera poses as reference frames. In Sec. 3 of the paper, we define $K$ to be the number of source views in our model. For the inference process, we select $K$ nearby optimized frames $\mathcal{I} = \{I_k\}_{k=1}^{K}$ as references for rendering pixels and optimizing the camera pose of the target frame. Empirically, we find that varying the stride of reference frame selection from 1 to 4 does not have a significant impact on the reconstruction results. To improve the network's capability to generalize across different view densities during the training process, we adopt the frame selection strategy from IBRNet [43]. We create a pool of $nK$ nearby views by randomly selecting $n$ from a uniform distribution of $[1, 3]$. Next, we randomly select $K$ views from the pool to be the reference views.

**Tracking Details.** We use SuperPoint [7] to detect interest points in RGB images and SuperGlue [29] to establish point correspondence between reference images and the target image, with estimated confidence scores for each point match. We then select the top $M_t$ pixel pairs between the reference images and the target image with the highest confidence scores. To ensure a fair comparison with baseline models, we set $M_t$ to 200 for the Replica Dataset [32], 1,000 for the ScanNet Dataset [6], and 2,000 for the TUM-RGBD Dataset [33]. In practice, a simpler approach would be to set a threshold for the confidence level of the matched point pairs.

**Dense Mesh Reconstruction.** Our model utilizes a neural implicit representation of the scene, and to reconstruct the 3D mesh, we store the $xyz$ points from the rendered $xyz$-maps. We set the default pixel stride for rendering to 5 and the default temporal frame stride to 10 (render every other 10 frames while tracking the pose of every frame). Tab. 6 compares the performance with various rendering pixel stride and temporal stride values. Once we have visited all the sequence frames and rendered novel views, we merge all stored $xyz$ points into a point cloud, remove outliers, and prune points that are too close. We can then calculate the rendered mesh structure from the rendered maps using Poisson surface reconstruction [15]. For scenes in NICE-SLAM Dataset, to generate mesh structure, we sample 8,160 pixels in every frame and render 200 frames per scene. The rendered $xyz$-map and RGB-map are saved as points' positions and colors. We adopt Poisson surface reconstruction [15] to generate 3D mesh from point clouds. We use `open3d.geometry.TriangleMesh` package in the implementation. Specifically, we determine the depth of the octree subdivision to be 8, and we assign the average distance of all points to be the size of the spherical neighborhoods used by the Poisson reconstruction algorithm.

| Point stride | Acc. ↓ | Comp. ↓ | Comp. Ratio ↑ |
|---|---|---|---|
| 5 | 1.83 | 2.02 | 93.79 |
| 10 | 1.90 | 2.09 | 93.49 |
| 15 | 2.24 | 2.31 | 93.03 |
| 20 | 2.31 | 2.49 | 92.36 |
| Temporal stride | Acc. ↓ | Comp. ↓ | Comp. Ratio ↑ |
| 5 | 1.80 | 1.99 | 93.81 |
| 10 | 1.83 | 2.02 | 93.79 |
| 20 | 2.27 | 2.44 | 93.03 |

Table 6. **3D reconstruction results with different points rendering strides and temporal skip strides on Replica Dataset [32].**

| Iterations | iMAP | NICE-SLAM | Ours |
|---|---|---|---|
| 10 | 3.42 | 2.05 | 0.83 |
| 15 | 2.72 | 1.69 | 0.62 |
| 20 | 2.23 | 1.21 | 0.59 |

Table 7. **Average tracking results with different optimization iterations.** ATE RMSE [cm] is used as the evaluation metric.

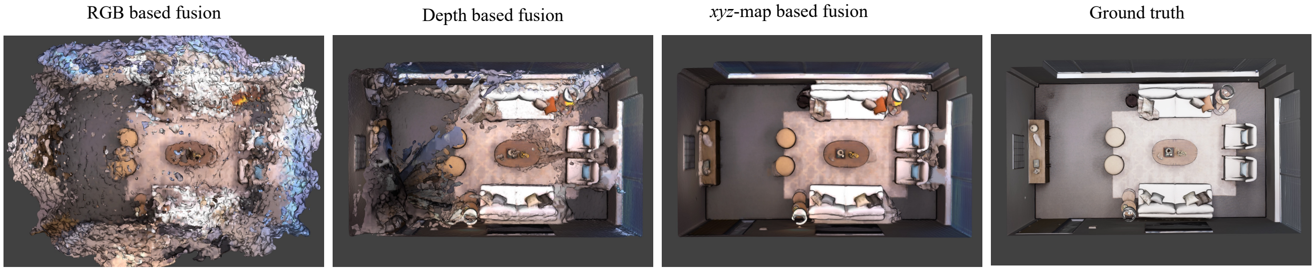
## B. Additional Results

**More Quantitative Results on the Replica Dataset [32].** In Tab. 8, we present comprehensive reconstruction results for each scene in the Replica Dataset [32], with quantitative values for iMAP and NICE-SLAM obtained from [53]. Our method demonstrates comparable performance with other per-scene optimization methods while outperforming them in terms of 2D metrics. In particular, our model generalizes to novel scenes unseen during training and reconstructs accurate geometric details, without any further finetuning of the NeRF model.

***xyz*-maps versus Other Variants for Fusion.** Fig. 6 shows reconstruction results on the Replica Dataset [32] using different fusion method. As described in Sec. 3.2 in the main paper, the most straightforward fusion method is to directly apply IBRNet and reconstruct the surface by integration over the density field. We denote this method as "RGB based fusion". Further introducing a depth fusion branch appears to be a more reasonable alternative, denoted as "Depth based fusion". As can be seen, both the

| | Metrics | room-0 | room-1 | room-2 | office-0 | office-1 | office-2 | office-3 | office-4 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| iMAP [35] | Depth L1↓ | 3.41 | 3.04 | 4.04 | 3.77 | 4.30 | 7.63 | 4.93 | 4.38 | 4.39 |
| | Acc. ↓ | 4.13 | 3.95 | 4.11 | 5.02 | 7.24 | 5.36 | 4.12 | 4.26 | 4.77 |
| | Comp. ↓ | 4.84 | 4.85 | 4.74 | 4.03 | 4.72 | 6.06 | 5.14 | 5.78 | 5.02 |
| | Comp. Ratio ↑ | 78.20 | 77.41 | 81.64 | 79.64 | 76.80 | 65.16 | 75.90 | 69.31 | 75.51 |
| DI-Fusion [13] | Depth L1 ↓ | 7.94 | 69.21 | 28.47 | 7.77 | 7.67 | 12.28 | 9.91 | 10.44 | 19.21 |
| | Acc. ↓ | 2.48 | 39.36 | 21.82 | 58.02 | 2.76 | 2.33 | 2.71 | 2.27 | 19.40 |
| | Comp. ↓ | 4.42 | 31.97 | 14.55 | 4.11 | 2.49 | 5.50 | 4.41 | 6.02 | 9.19 |
| | Comp. Ratio ↑ | 90.08 | 41.98 | 56.75 | 89.79 | 89.60 | 86.14 | 85.70 | 84.43 | 78.06 |
| NICE-SLAM [53] | Depth L1 ↓ | 1.88 | 1.67 | 2.23 | 1.75 | 2.02 | 4.82 | 3.50 | 2.01 | 2.49 |
| | Acc. ↓ | 2.33 | 2.26 | 2.27 | 2.01 | 2.16 | 3.10 | 2.88 | 2.36 | 2.42 |
| | Comp. ↓ | 2.57 | 2.24 | 2.64 | 1.86 | 2.13 | 3.12 | 3.31 | 3.34 | 2.65 |
| | Comp. Ratio ↑ | 91.53 | 92.80 | 89.94 | 94.17 | 94.07 | 86.92 | 84.93 | 87.05 | 90.30 |
| Ours | Depth L1 ↓ | 0.82 | 1.45 | 1.65 | 1.61 | 0.99 | 1.78 | 2.57 | 1.37 | **1.53** |
| | Acc. ↓ | 1.51 | 1.67 | 1.82 | 1.84 | 1.66 | 1.99 | 2.47 | 1.69 | **1.83** |
| | Comp. ↓ | 1.30 | 1.81 | 2.23 | 1.97 | 1.73 | 2.49 | 2.88 | 1.75 | **2.02** |
| | Comp. Ratio ↑ | 96.08 | 95.96 | 93.14 | 93.29 | 94.14 | 92.11 | 91.57 | 94.08 | **93.79** |

Table 8. **Reconstruction results on different scenes in Replica Datasets.**


RGB based fusion

Depth based fusion

*xyz*-map based fusion

Ground truth

Figure 6. **Reconstruction results with different fusion methods.** Our $xyz$-map based fusion substantially outperforms the other baselines based on IBRNet.

results by RGB based fusion and depth based fusion are far from satisfactory. It is unserprising that the RGB based fusion struggles to achieve good performance due to the lack of geometry clues in the rendering process. For the depth based fusion, the geometry exhibited inconsistencies and inaccuracies. We hypothesize that it is due to the fact that the depth maps are represented in view-dependent camera coordinates, resulting in the multi-view inconsistency. In contrast, our *xyz*-map based fusion leads to superior reconstruction results, which inherently considers the multi-view consistency by design. These results further highlight the strength of our IBD-SLAM framework.

**Geometry Completion.** As depicted in Fig. 7, our model possesses the capability to complete holes and missing regions with novel view synthesis. Some regions of the scene might still not be seen halfway through the sequence optimization. To generate novel views, we slightly perturb the currently optimized camera poses. Our model can make a reasonable estimation of these unseen areas and fill in the missed regions. In Fig. 7, our model successfully fills in the geometry holes at frames 750, 1,000, and 1,500, which is a testament to its novel view synthesis capacity. Since our model renders novel views of a scene, it produces a better coverage of the scene geometry.

**Point Rendering Stride and Temporal Skip Stride.** In Tab. 6, We evaluate the effects of two factors that affect the reconstruction performance, namely point rendering stride and the temporal skip stride. A small point rendering stride allows for more accurate reconstruction, while at the cost of more time and memory consumption. We simply adopt the default stride by IBRNet. We also evaluate different temporal skip strides on Replica Dataset [32]. We can see that skip strides of 5 and 10 lead to similar results, while when the stride is set to 20, all metrics deteriorate rapidly. We choose the stride of 10 for the performance and efficiency trade-off.

**Ablation with Different Numbers of Tracking Iterations.** We vary the number of tracking iterations and evaluate the

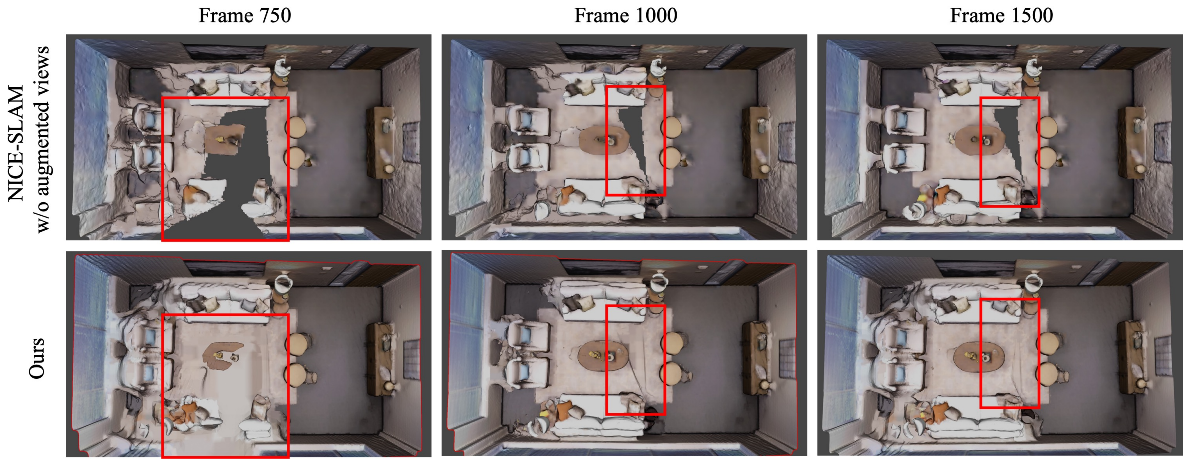


Figure 7. **The geometry completion ability of our model.** Our method can faithfully fill in the missing gaps of the scene.

final results by calculating the ATE RMSE [cm]. The figures in the table represent the average results of the Replica Dataset scenes. Our method requires fewer iterations to achieve a reasonably accurate camera pose, as illustrated in Tab. 7. Therefore, in our experiments, we use 15 steps for camera pose optimization per tracking frame.

**More Reconstruction Results.** The reconstruction results comparison of Co-SLAM, ESLAM, and our method are shown in Fig. 8 and Fig. 9. The reconstruction results of ScanNet Dataset [6] `scene0000_00` are shown in Fig. 11. More reconstruction results of Replica Dataset [32] are also shown in Fig. 10. As can be seen, our method can faithfully reconstruct the scenes with fine details, despite not requiring per-scene retraining.

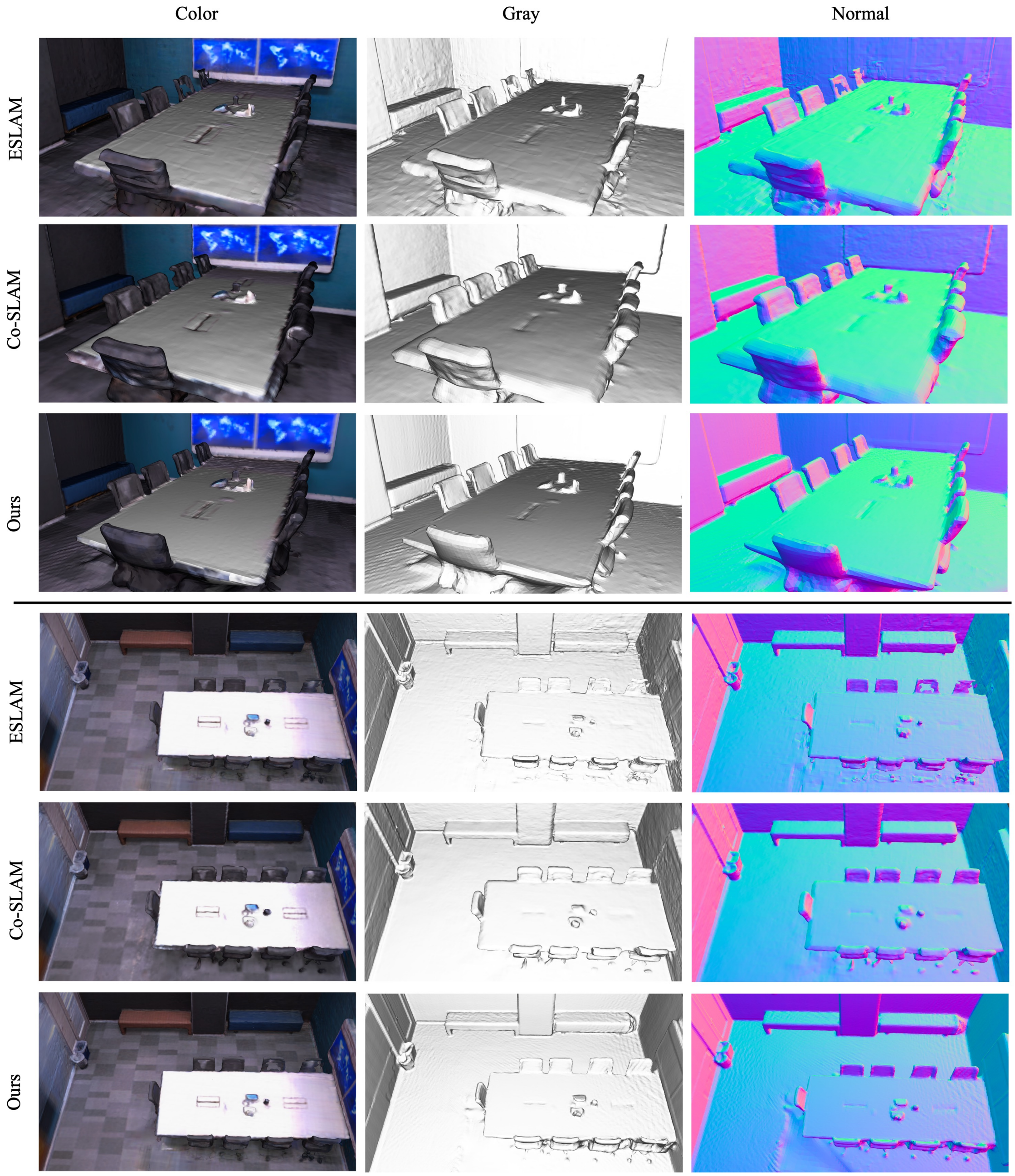


Figure 8. **Reconstructioin results comparison with Co-SLAM [42] and ESLAM [14].** We present reconstruction results on office4 of Replica Dataset [32].

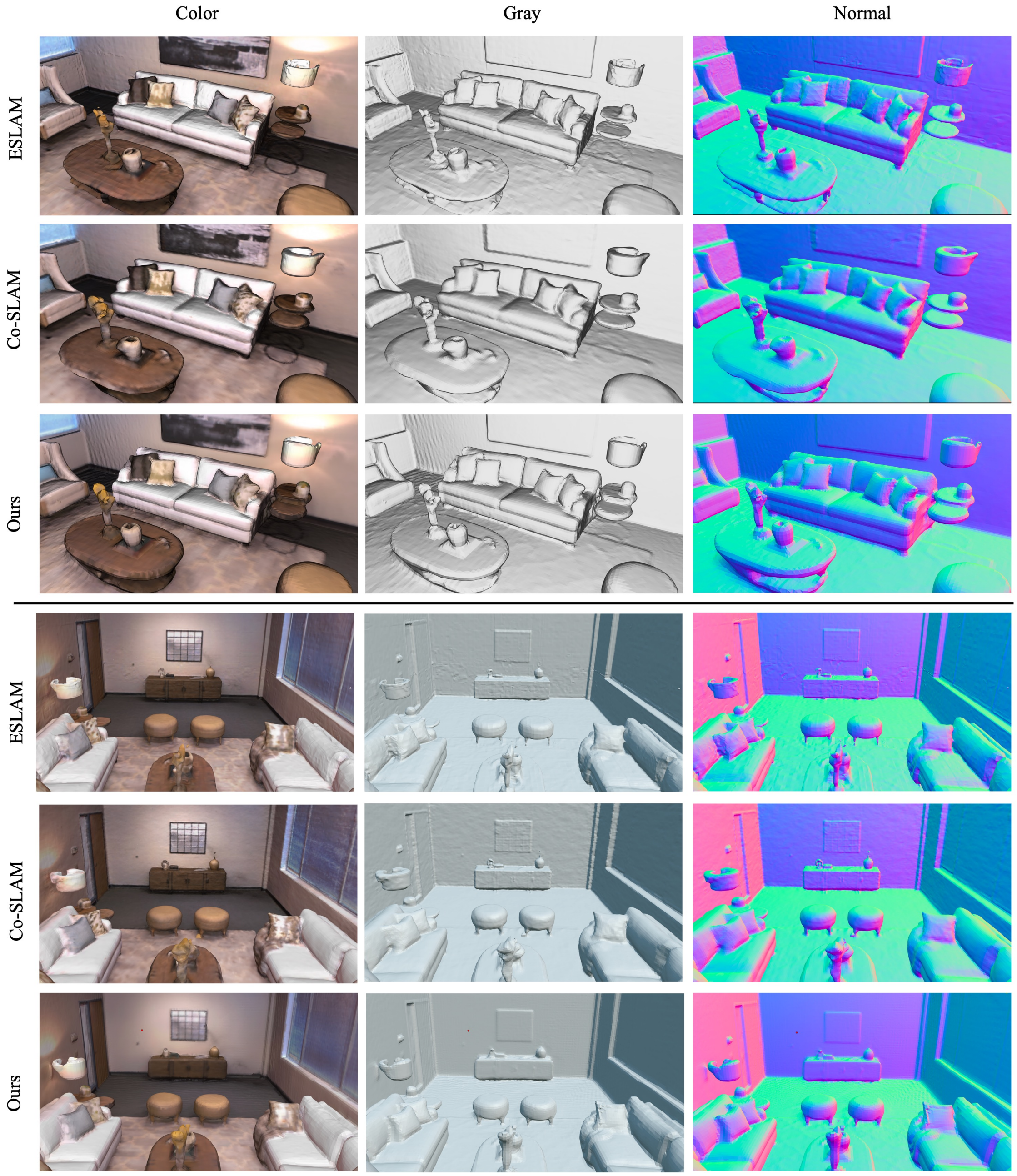


Figure 9. **Reconstructioin results comparison with Co-SLAM [42] and ESLAM [14].** We present reconstruction results on room0 of Replica Dataset [32].

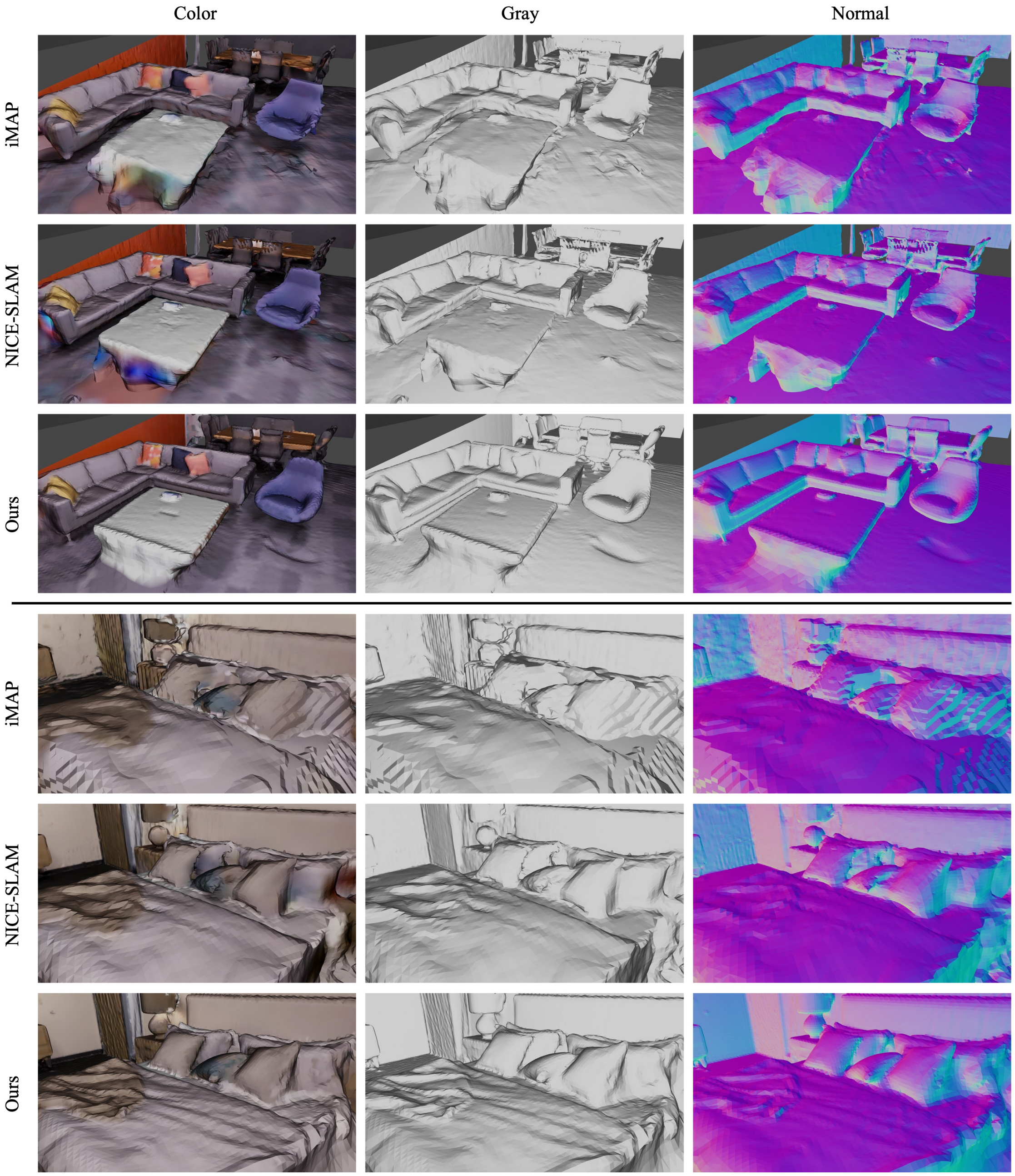


Figure 10. **More detailed reconstruction results on Replica Dataste [32].** We present reconstruction results on room1 (top three rows) and office (bottom three rows).

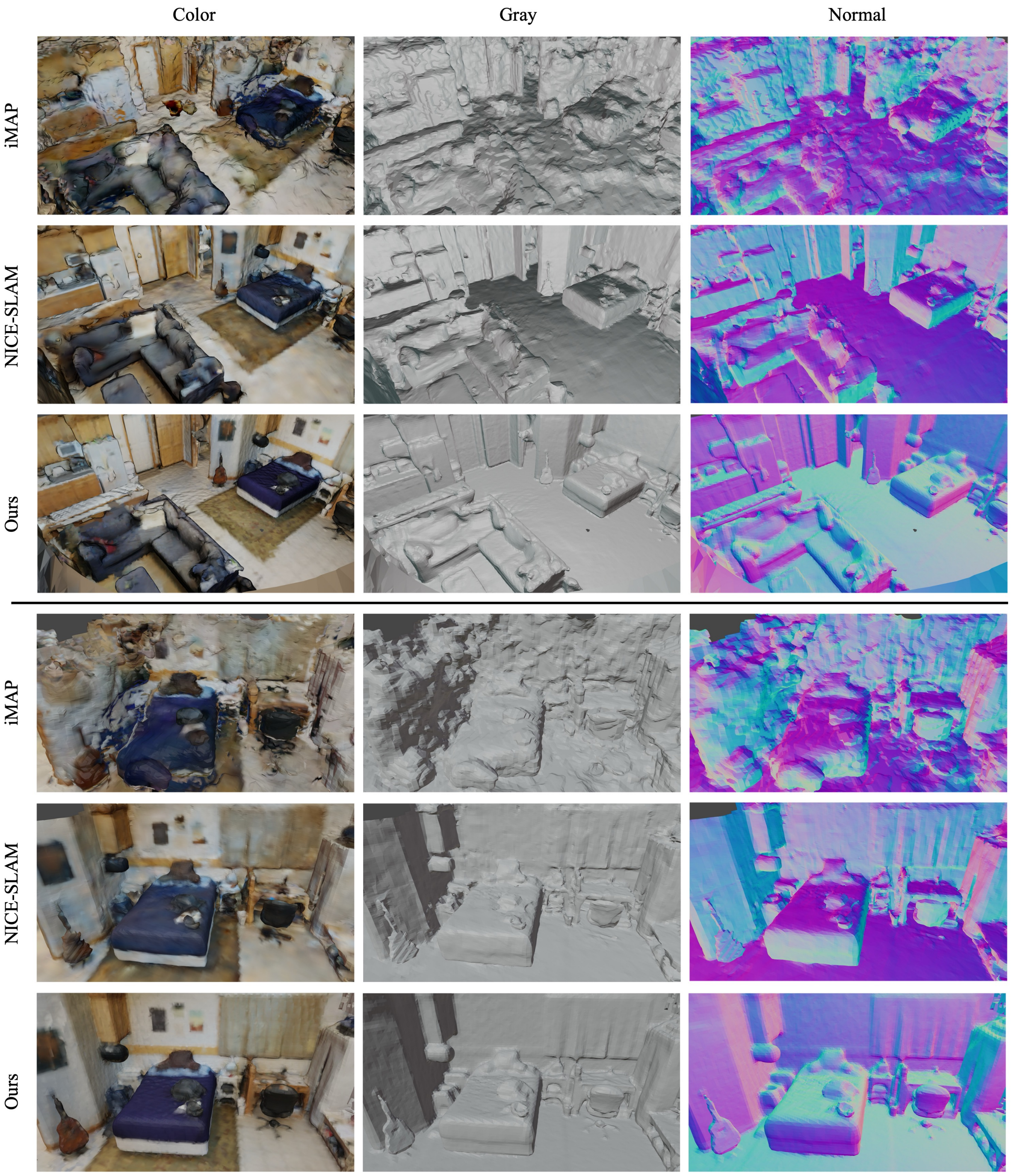


Figure 11. **3D reconstruction results on ScanNet Dataset [6].** We present reconstruction results on $scene0000\_00$.