IndoorGS: Geometric Cues Guided Gaussian Splatting for Indoor Scene Reconstruction

Supplementary Material

In this supplementary material, we provide implementation details in Appendix A, out-of-distribution novel view synthesis experiment in Appendix B, and further per-scene quantitative and qualitative results in Appendix C and Appendix D respectively.

A. Implementation details

A.1. Extracting Plane-like Cues

Algorithm 1 Pseudo-code for Extracting Plane-like Cues

$D_{\text{mono}} = \text{Mono-model}(I)$ ▷ Estimate mono-depth
$D_{\text{sparse}} = D_{\text{SfM}} \cup D_{\text{Line}}$
$\hat{k} = \arg\min_k \sum \|k \cdot D_{\text{mono}} - D_{\text{sparse}}\|_2^2$ ▷ Regression
$D_{\text{adjust}} = \hat{k} \cdot D_{\text{mono}}$ ▷ Depth align
$A_i = SAM(I)$ ▷ Segment image
for each A_i **do** ▷ Detect textureless area
  if SIFT point density $< D_{\text{thresh}}$ **then**
    Mark A_i as weakly textured
  end if
end for
for each textureless area A_i **do**
  $P_i = R^{-1}(K^{-1} D_{\text{adjust}} A_i - t)$ ▷ Back project
  $L_{2\text{d}} = Lines$ in $dilate(A_i, 3)$
  $L_{3\text{d}} = $ 3d lines corresponded to $L_{2\text{d}}$
  for each endpoint p_j of 3D line $L_{3\text{d}}$ **do**
    if $d_j = \text{NN}(p_j, P_i) > d_{\text{thresh}}$ **then**
      Discard P_i
    else
      Retain P_i
    end if
  end for
end for
$P = \sum(\text{Retained } P_i)$ ▷ Merge results
Plane cues $=$ downsample(P) ▷ Output

Algorithm 1 employs 2 key hyperparameters: D_{thresh}, the SIFT point density threshold, determined as 0.001 through multi-image density analysis and texture strength classification; and d_{thresh}, the endpoint-to-point cloud distance threshold, defined as triple the average point cloud distance.

A.2. Geometric cues input

As shown in Fig. 6, IndoorGS utilizes denser geometric cues than sparse 3D SfM points as the initial input data. The density is controlled during optimization using different ADC (Adaptive Density Control) strategies tailored to various data types. To achieve this goal, we introduce a type attribute to each Gaussian primitive to record their specific types: type=0 for line Gaussians, type=1 for SfM Gaussians, and type=2 for planar-like Gaussians. During the densification process, newly added Gaussians, i.e. those resulting from cloning and splitting operations, inherit the type attribute of their parent Gaussian.

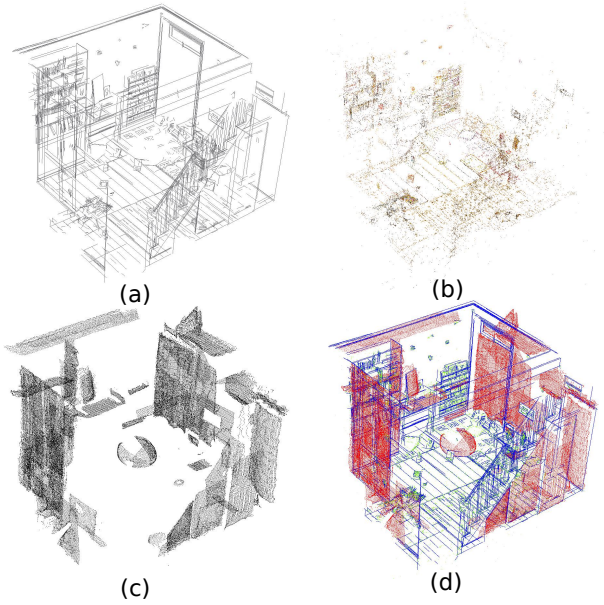


Figure 6. **Three types of geometric cues as initial inputs on the "playroom" scene of the Deep Blending dataset. (a)** geometric cues from 3D line segments. **(b)** geometric cues from SfM. **(c)** geometric cues from planes. **(d)** All input cues, green points represent SfM cues, blue points represent line cues, and red points represent plane-like cues.

A.3. SOR filter

Section 4.1 uses the Statistical Outlier Removal (SOR) filter to clean the sparse SfM point cloud by identifying and removing noise or outliers. It calculates the average distance between each point and its K-nearest neighbors, then removes points whose average distances deviate from the mean. Specifically, a point p_i is considered an outlier if its mean distance $\bar{d}_i$ exceeds a threshold defined by the global mean μ and a multiple of the standard deviation σ:

$$\bar{d}_i > \mu + \beta\sigma \tag{15}$$

where β is a threshold parameter that determines how many standard deviations above the mean a point's distance must

be to be classified as an outlier. The mean distance for point p_i is given by $\bar{d}_i = \frac{1}{K}\sum_{j=1}^{K}\|p_i - p_j\|$. The global mean distance across the entire point cloud is expressed as $\mu = \frac{1}{N}\sum_{i=1}^{N}\bar{d}_i$, and the standard deviation of the distances is given by $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\bar{d}_i - \mu\right)^2}$, N denotes the number of points in the point cloud. This process improves the overall quality and reliability of the point cloud for further processing or analysis. We set K=5 and β=1.3 in our experiments.

Figure 7. **A comparison of the SfM point cloud before and after filtering on the "playroom" scene of the Deep Blending dataset.** The point cloud on the left represents the pre-filtered state with 37,005 points, and the point cloud on the right represents the post-filtered result with 36,850 points. In comparison, the right side has fewer noisy points than the left side.

A.4. Dataset

Replica. The Replica dataset, being synthetic, includes living room and office scenarios. We selected eight scenarios for our experiments: "Room0-2" and "Office0-4". The dataset was obtained from NICE-SLAM [52] to simplify the reconstruction process. Specifically, we sampled one image every 10 frames from the 2000 images available per scene, resulting in 200 images per scene for our experiments.

ScanNet++. For the ScanNet++ dataset, we used half-resolution images for reconstruction due to the high image resolution and excluded any images labeled as blurry. We used DSLR camera sequences with COLMAP-registered poses for four scenes, including "a5c013435", "3f1e1610de", "66c98f4a9b" and "88cf747085", which contain 105, 185, 67, and 182 registered images.

Deep Blending. The Deep Blending dataset is a real-world dataset that contains high-resolution RGB images captured from multiple views of a real-world scene, along with precise camera parameters. We use two scenes, namely "Drjohnson" and "Playroom", which contain 263 and 225 high-resolution photos, respectively.

A.5. Mesh evaluation metrics

Tab. 4 defines the mesh evaluation metrics to compare the predicted mesh with the ground truth mesh for the Replica and ScanNet++ datasets. A threshold of 5 cm is used for calculating precision, recall, and F1-score. Additionally, the mesh quality assessment is limited to the area within the visual range of the training camera views. Specifically, we crop the ground truth (GT) mesh by determining the visibility of its vertices and use the cropped mesh as our ground truth reference.

Metric	Definition
Accuracy	$\frac{1}{\lvert P\rvert}\sum_{\boldsymbol{p}\in P}\left(\min_{\boldsymbol{p}^*\in P^*}\|\boldsymbol{p}-\boldsymbol{p}^*\|_1\right)$
Completion	$\frac{1}{\lvert P^*\rvert}\sum_{\boldsymbol{p}^*\in P^*}\left(\min_{\boldsymbol{p}\in P}\|\boldsymbol{p}-\boldsymbol{p}^*\|_1\right)$
Chamfer-L_1	$\frac{\text{Accuracy + Completion}}{2}$
Precision	$\frac{1}{\lvert P\rvert}\sum_{\boldsymbol{p}\in P}\left(\min_{\boldsymbol{p}^*\in P^*}\|\boldsymbol{p}-\boldsymbol{p}^*\|_1 < 0.05\right)$
Recall	$\frac{1}{\lvert P^*\rvert}\sum_{\boldsymbol{p}^*\in P^*}\left(\min_{\boldsymbol{p}\in P}\|\boldsymbol{p}-\boldsymbol{p}^*\|_1 < 0.05\right)$
F1-score	$\frac{2\cdot \text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$
Normal-Accuracy	$\frac{1}{\lvert P\rvert}\sum_{\boldsymbol{p}\in P}\left(\boldsymbol{n}_{\boldsymbol{p}}^T\boldsymbol{n}_{\boldsymbol{p}^*}\right)$ s.t. $\boldsymbol{p}^* = \underset{p^*\in P^*}{\text{argmin}}\|\boldsymbol{p}-\boldsymbol{p}^*\|_1$
Normal-Completion	$\frac{1}{\lvert P^*\rvert}\sum_{\boldsymbol{p}^*\in P^*}\left(\boldsymbol{n}_{\boldsymbol{p}}^T\boldsymbol{n}_{\boldsymbol{p}^*}\right)$ s.t. $\boldsymbol{p} = \underset{p\in P}{\text{argmin}}\|\boldsymbol{p}-\boldsymbol{p}^*\|_1$
Normal-Consistency	$\frac{\text{Normal-Acc+Normal-Comp}}{2}$

Table 4. **Evaluation Metrics.** We show the evaluation metrics with their definitions that we use to measure reconstruction quality. P and P^* are the point clouds sampled from the predicted and the ground truth mesh. $\boldsymbol{n_p}$ is the normal vector at point $\boldsymbol{p}$.

B. Out-of-distribution NVS experiment

We follow the out-of-distribution novel view synthesis (OOD-NVS) definition of SplatFormer [7]. Test views significantly differ from input views in the out-of-distribution novel view synthesis(OOD-NVS). This scenario contrasts with in-distribution NVS, where test views interpolate between densely captured input views. According to SplatFormer, the geometrically consistent scene representation demonstrates a significant advantage in the OOD-NVS task. Our result is also consistent with the underlying geometry. To highlight the effectiveness of our approach for the NVS task, we propose a simple method for extracting an out-of-distribution test set. Specifically, we compute the nearest-neighbor distances between the centers of all cameras and sort them in descending order. The test set is constructed using data corresponding to the top one-eighth of cameras with the largest distances.

We then collected an indoor dataset consisting of 163 images casually captured with a cellphone within a single room. After processing with COLMAP to calculate intrinsic and extrinsic camera parameters and correct image distortion, each image was resized to 1195 × 896 pixels.

We extracted OOD views as test views using the above-mentioned method and then extracted geometric cues using our proposed geometric cue extraction approach (Section 4.1). As illustrated in Fig. 8, our geometric cue inputs on the right are significantly denser than the 3DGS input data

Method	Metric	R0	R1	R2	OFF0	OFF1	OFF2	OFF3	OFF4	Avg.
3DGS	PSNR↑	37.45	38.62	39.82	43.06	41.91	38.12	37.75	39.45	39.52
	SSIM↑	0.964	0.966	0.969	0.980	0.966	0.967	0.964	0.966	0.968
	LPIPS↓	0.097	0.115	0.123	0.079	0.179	0.139	0.114	0.122	0.121
GOF	PSNR↑	39.24	40.06	41.33	44.28	43.67	39.38	39.37	40.51	40.98
	SSIM↑	0.978	0.978	0.981	0.986	0.979	0.981	0.979	0.978	0.980
	LPIPS↓	0.060	0.074	0.072	0.055	0.106	0.072	0.071	0.081	0.074
RaDeGS	PSNR↑	39.24	40.30	41.48	44.41	43.60	39.34	39.35	40.61	41.04
	SSIM↑	0.978	0.978	0.981	0.986	0.979	0.981	0.978	0.978	0.980
	LPIPS↓	0.061	0.073	0.073	0.054	0.107	0.073	0.072	0.082	0.074
PGSR	PSNR↑	38.15	39.31	40.32	43.70	43.06	38.28	37.91	39.46	40.02
	SSIM↑	0.971	0.973	0.974	0.984	0.976	0.973	0.970	0.972	0.974
	LPIPS↓	0.087	0.100	0.112	0.065	0.127	0.109	0.100	0.105	0.101
Ours	PSNR↑	39.12	40.13	41.28	44.47	44.18	39.33	39.29	41.09	41.11
	SSIM↑	0.978	0.978	0.980	0.987	0.982	0.980	0.978	0.979	0.980
	LPIPS↓	0.058	0.078	0.075	0.050	0.089	0.081	0.071	0.081	0.073

Table 5. Quantitative comparison results of rendering quality for novel view synthesis on the Replica dataset.

Method	Metric	R0	R1	R2	OFF0	OFF1	OFF2	OFF3	OFF4	Avg.
GOF	C-L1↓	0.069	0.084	0.117	0.067	0.107	0.083	0.068	0.092	0.086
	F1↑	0.654	0.566	0.493	0.666	0.425	0.577	0.668	0.562	0.576
	NC↑	0.838	0.796	0.789	0.807	0.668	0.817	0.842	0.818	0.797
RaDeGS	C-L1↓	0.068	0.080	0.093	0.051	0.132	0.067	0.058	0.080	0.079
	F1↑	0.666	0.561	0.533	0.709	0.391	0.647	0.690	0.603	0.600
	NC↑	0.885	0.867	0.864	0.882	0.750	0.879	0.887	0.883	0.862
PGSR	C-L1↓	0.045	0.046	0.083	0.030	0.062	0.045	0.034	0.045	0.049
	F1↑	0.850	0.818	0.724	0.886	0.734	0.815	0.873	0.844	0.818
	NC↑	0.903	0.885	0.862	0.923	0.875	0.892	0.915	0.908	0.895
Ours	C-L1↓	0.032	0.029	0.041	0.023	0.044	0.046	0.033	0.042	0.036
	F1↑	0.893	0.902	0.774	0.907	0.739	0.715	0.866	0.803	0.825
	NC↑	0.956	0.947	0.946	0.949	0.918	0.932	0.934	0.940	0.940

Table 6. Quantitative comparison results of Chamfer distance, F-Score, and Normal Consistency for reconstruction on the Replica dataset.

Method	Metric	0a5c	3f1e	66c9	88cf	Avg.
3DGS	PSNR↑	29.37	31.96	25.26	31.23	29.46
	SSIM↑	0.933	0.944	0.864	0.930	0.918
	LPIPS↓	0.143	0.147	0.175	0.143	0.152
GOF	PSNR↑	29.68	32.47	25.94	31.17	29.82
	SSIM↑	0.932	0.950	0.864	0.930	0.919
	LPIPS↓	0.141	0.146	0.179	0.146	0.153
RaDeGS	PSNR↑	30.52	32.19	26.50	31.10	30.08
	SSIM↑	0.938	0.943	0.871	0.931	0.921
	LPIPS↓	0.131	0.150	0.165	0.141	0.147
PGSR	PSNR↑	30.66	32.16	25.97	31.17	29.99
	SSIM↑	0.936	0.950	0.863	0.932	0.920
	LPIPS↓	0.142	0.147	0.182	0.146	0.154
Ours	PSNR↑	31.17	32.05	27.24	31.35	30.45
	SSIM↑	0.941	0.955	0.880	0.935	0.928
	LPIPS↓	0.125	0.126	0.140	0.134	0.131

Table 7. Quantitative comparison results of rendering quality for novel view synthesis on the Scannet++ dataset. Row one, '0a5c, 3f1e, 66c9, 88cf', is an abbreviation for the full hexadecimal strings '0a5c013435, 3f1e1610de, 66c98f4a9b, 88cf747085'.

Method	Metric	Drjohnson	Playroom	Avg.
3DGS	PSNR↑	28.77	30.04	29.41
	SSIM↑	0.899	0.906	0.903
	LPIPS↓	0.244	0.241	0.243
GOF	PSNR↑	28.24	30.17	29.21
	SSIM↑	0.897	0.910	0.904
	LPIPS↓	0.253	0.239	0.246
RaDeGS	PSNR↑	28.81	30.1	29.46
	SSIM↑	0.902	0.911	0.907
	LPIPS↓	0.246	0.240	0.243
PGSR	PSNR↑	28.18	30.25	29.22
	SSIM↑	0.877	0.909	0.893
	LPIPS↓	0.264	0.245	0.255
Ours	PSNR↑	29.51	30.61	30.06
	SSIM↑	0.907	0.913	0.910
	LPIPS↓	0.235	0.235	0.235

Table 8. Quantitative comparison results of rendering quality for novel view synthesis on the Deep Blending dataset

shown on the left. We evaluated our method and 3DGS on this dataset. The 3DGS reconstruction achieved a PSNR of only 13.47 dB, essentially considered a failure. In contrast, our method achieved a PSNR of 27.93 dB on test views, corresponding to significantly better performance as shown in Fig. 9.

This experiment highlights the robustness of our method across different datasets and underscores its comparative advantage in handling OOD test views.

Method	Metric	0a5c	3f1e	66c9	88cf	Avg.
GOF	C-L1↓	0.085	0.139	0.133	0.061	0.105
	F1↑	0.487	0.387	0.424	0.618	0.479
	NC↑	0.669	0.702	0.680	0.756	0.702
RaDeGS	C-L1↓	0.075	0.127	0.120	0.061	0.096
	F1↑	0.533	0.389	0.446	0.620	0.497
	NC↑	0.723	0.768	0.746	0.806	0.761
PGSR	C-L1↓	0.065	0.099	0.100	0.051	0.079
	F1↑	0.651	0.561	0.585	0.703	0.625
	NC↑	0.727	0.810	0.751	0.825	0.778
Ours	C-L1↓	0.037	0.058	0.042	0.024	0.040
	F1↑	0.777	0.686	0.720	0.922	0.776
	NC↑	0.817	0.890	0.867	0.933	0.877

Table 9. Quantitative comparison results of Chamfer distance, F-Score, and Normal Consistency for reconstruction on the Scannet++ dataset.

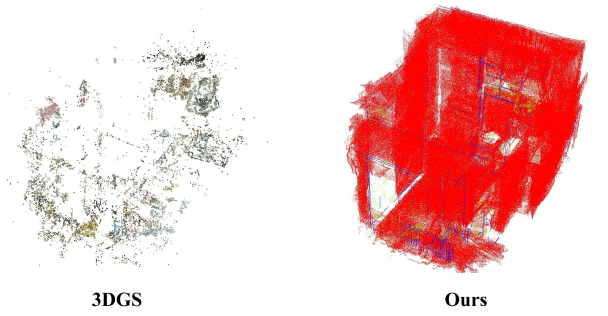


Figure 8. **Different initialization input data.**

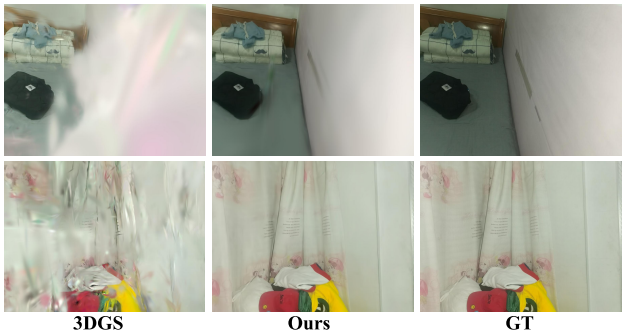


Figure 9. **Qualitative comparison of test view results.**

C. Per-scene quantitative results

In this section, we present per-scene quantitative results. Tab. 5, Tab. 7, and Tab. 8 show the NVS comparison on the Replic, Scannet++, and Deep Blending datasets respectively, while Tab. 6 and Tab. 9 shows the mesh quality comparison on the Replic and Scannet++ datasets.

D. More qualitative results

More qualitative results are given in Fig. 10, Fig. 11, Fig. 12, and Fig. 13.

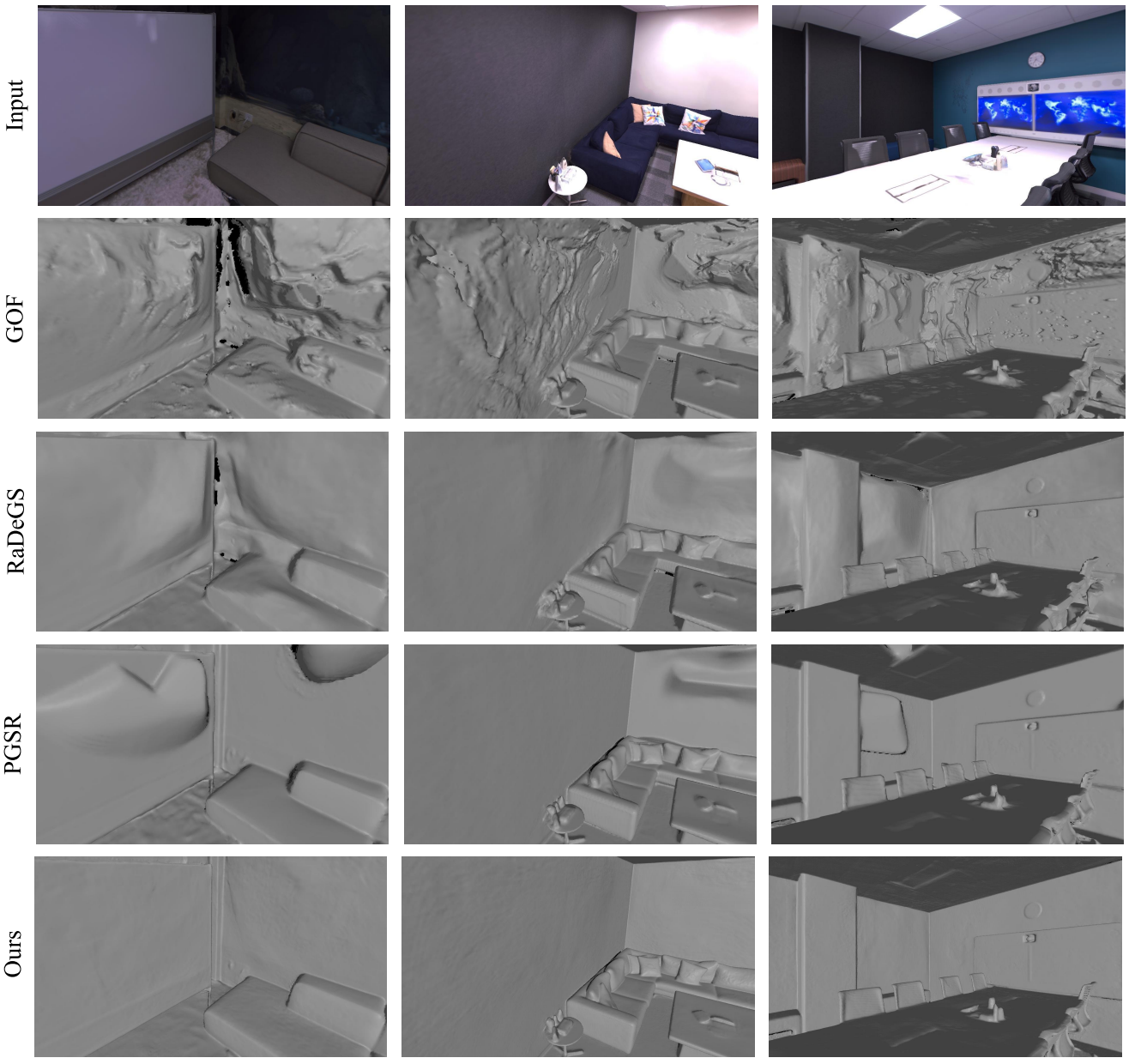


Figure 10. **Qualitative comparison on mesh reconstruction.** Comparison of baseline methods on sequences from the **Replic** dataset.

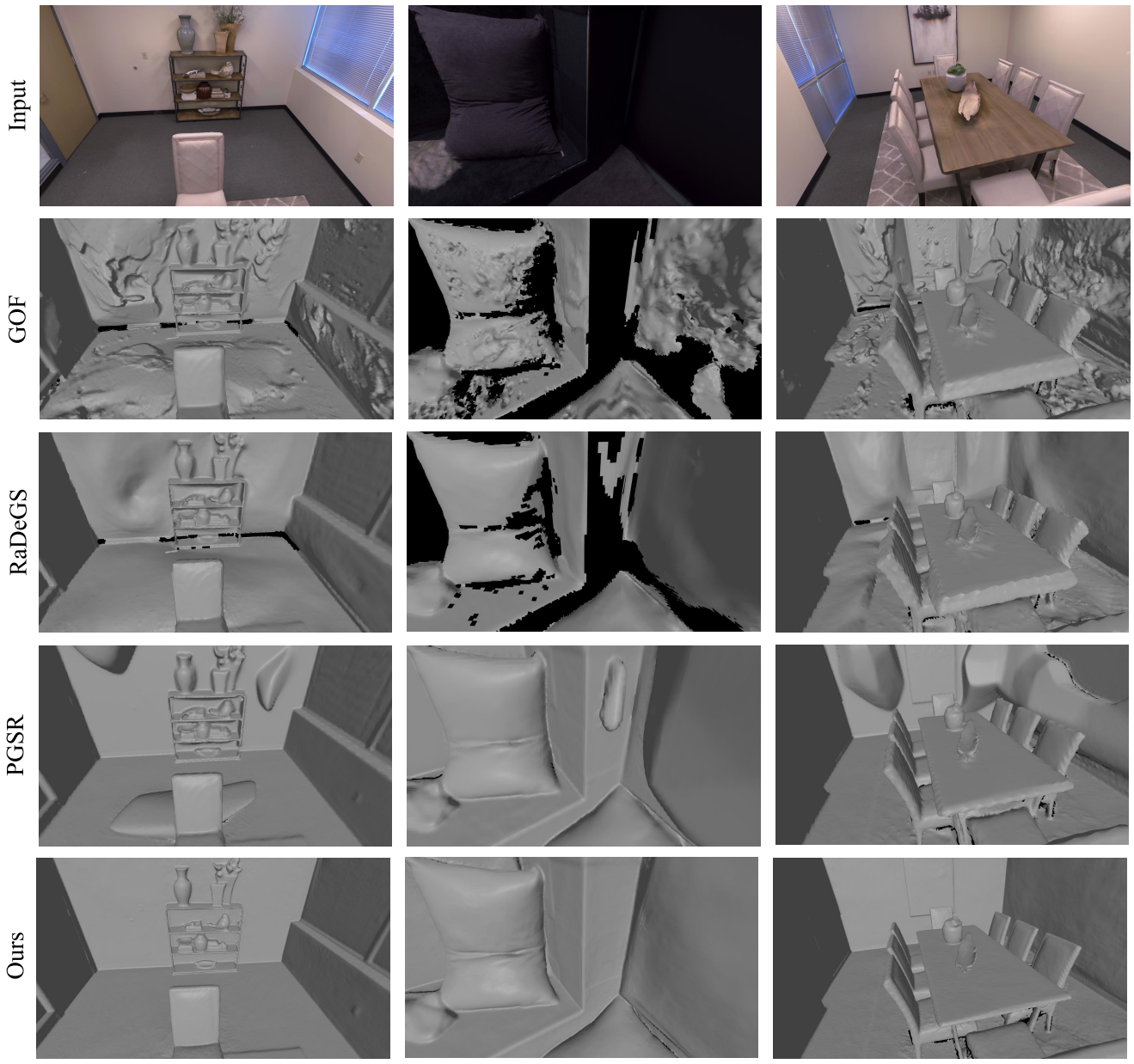
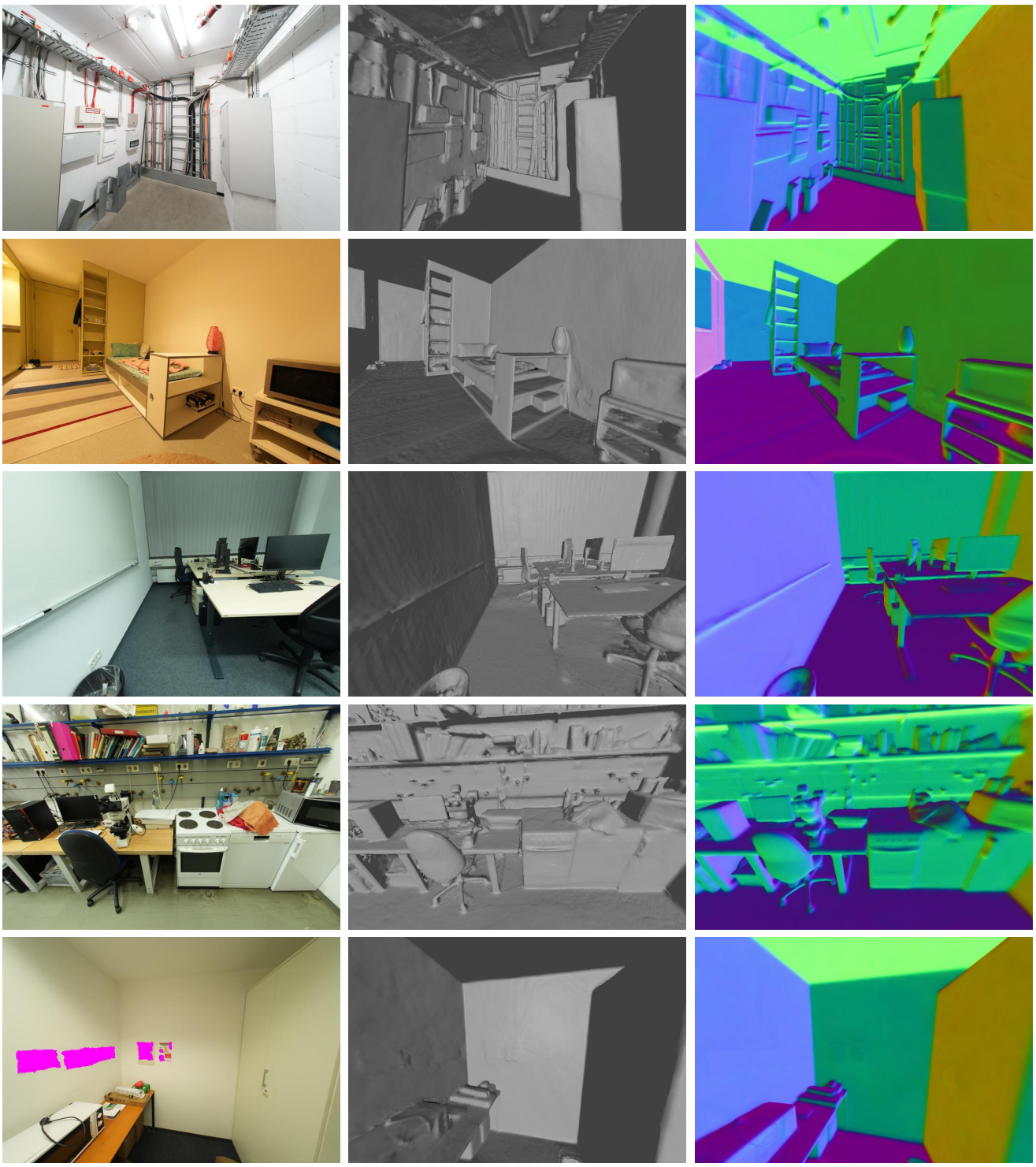
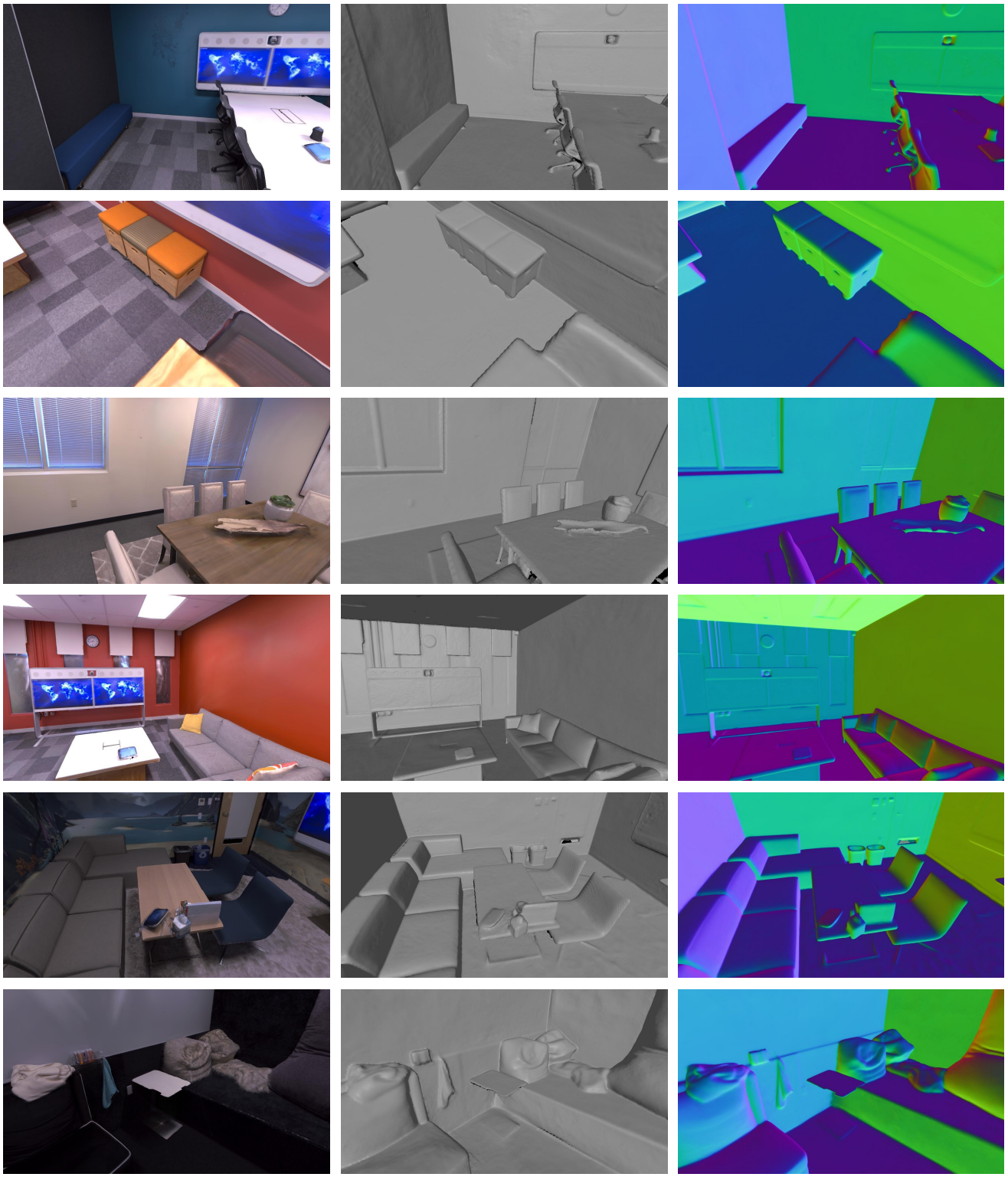


Figure 11. **Qualitative comparison on mesh reconstruction.** Comparison of baseline methods on sequences from the **Replic** dataset.



(a) Rendered RGB (b) Mesh (c) Mesh Normal

Figure 12. Additional experimental results of IndoorGS on the **ScanNet++** dataset.



(a) Rendered RGB (b) Mesh (c) Mesh Normal

Figure 13. IndoorGS enables **high-precision geometric reconstruction and novel view synthesis** for indoor datasets using a sequence of RGB images.