

A Appendix

A.1 Planar Indicator Initialization

As described in Sec. 3.3, we initialize the plane indicator using ceiling and floor points derived from semantically lifted SfM points or semantic Gaussian primitives. The SfM points are generated by triangulating posed images through 2D feature matching, establishing 2D-3D correspondences. Utilizing these correspondences, we aggregate semantic labels for each 3D point from 2D semantic maps across multiviews and apply a voting procedure to identify the most prevalent semantic label, including those for ceiling and floor. The Gaussian semantic lifting module, mentioned in Sec. 3.2, lifts 2D semantic maps to each Gaussian primitive, and each primitive contains a semantic probability of the wall, floor, ceiling, or other categories. Consequently, SfM points and Gaussian primitives are assigned structural semantic labels such as wall, floor, or ceiling, allowing us to extract ceiling and floor points.

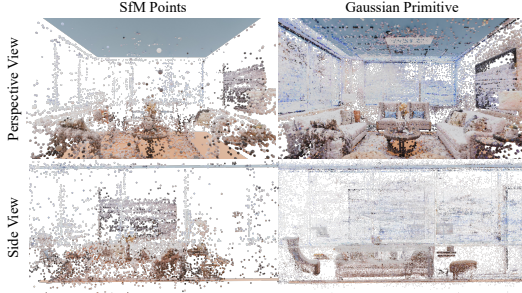


Figure A: **Plane Indicator Visualization.** We visualize the plane indicators derived from both the semantic lifted SfM points and the semantic Gaussian primitives from both the perspective and side views. In the visualization, the ceiling plane is colored in blue, while the floor plane is colored in orange.

Subsequently, we conduct plane fitting to identify the floor plane (\mathbf{n}_f, d_f) using RANSAC [4] applied to the extracted floor points. The normal vector \mathbf{n}_f is chosen as the gravity direction \mathbf{n}_g . The offset of the ceiling plane, d_c , is calculated based on the ceiling points and the gravity direction as follows:

$$d_c = \text{mean}_{\mathbf{p} \in \mathbf{P}_{\text{ceiling}}} (\mathbf{n}_g \cdot \mathbf{p}), \quad (1)$$

where $\mathbf{P}_{\text{ceiling}}$ represents the set of ceiling points. The plane indicator is initially determined using the semantic lifted points. If the angle deviation or the offset discrepancy surpasses a threshold, the plane indicator is reinitialized using semantic Gaussian primitives to minimize inaccuracies in textureless regions. Fig. A further illustrates plane indicators derived from both semantic lifted sparse points and semantic Gaussian primitives, demonstrating that both approaches can provide reliable structural priors.

A.2 Additional Implementation Details

Our implementation is based on PyTorch, utilizing customized surfel rasterization techniques for semantic learning. Parameters are optimized using the Adam optimizer. Most of the training learning rates are similar to those used in [8]. We set the hyperparameter \mathcal{K} to 10 for indoor scenes and 5 for urban scenes, with a voxel size of 0.01, and the feature dim is 32 in our sparse feature grid. For all scenes, the implicit-structured Gaussian is trained for 40,000 steps. Voxels grow between steps 1,500 and 20,000, provided the gradients of the Gaussians exceed $2e-4$ and are pruned if the opacities of all local Gaussians fall below 0.005. During training, we start our 3D global planar regularization from step 7000 and 2D local surface regularization from 20000. After completing training, surfaces are extracted using TSDF-Fusion [3], following the approach described in [6].

Table A: **Definition of metrics.** P and P^* are the 3D points from the predicted and the GT mesh.

Metric	Definition
Acc	$\text{mean}_{\mathbf{p} \in P} (\min_{\mathbf{p}^* \in P^*} \ \mathbf{p} - \mathbf{p}^*\)$
Comp	$\text{mean}_{\mathbf{p}^* \in P^*} (\min_{\mathbf{p} \in P} \ \mathbf{p} - \mathbf{p}^*\)$
CD	$\frac{\text{Acc} + \text{Comp}}{2}$
Prec	$\text{mean}_{\mathbf{p} \in P} (\min_{\mathbf{p}^* \in P^*} \ \mathbf{p} - \mathbf{p}^*\ < 0.05)$
Recall	$\text{mean}_{\mathbf{p}^* \in P^*} (\min_{\mathbf{p} \in P} \ \mathbf{p} - \mathbf{p}^*\ < 0.05)$
F1-score	$\frac{2 \times \text{Prec} \times \text{Recall}}{\text{Prec} + \text{Recall}}$

A.3 Additional Experimental Details

Similar to previous works for indoor scene reconstruction [14], we select four scenes in ScanNet [2], including *scene0050_00*, *scene0084_00*, *scene0580_00*, *scene0616_00* and seven scenes in Replica [10], *office0-office3*, *room0-room2*, and as for ScanNet++ [13], we select four scenes, *8b5caf3398*, *b20a261fdf*, *f34d532901*, *f6659a3107*.

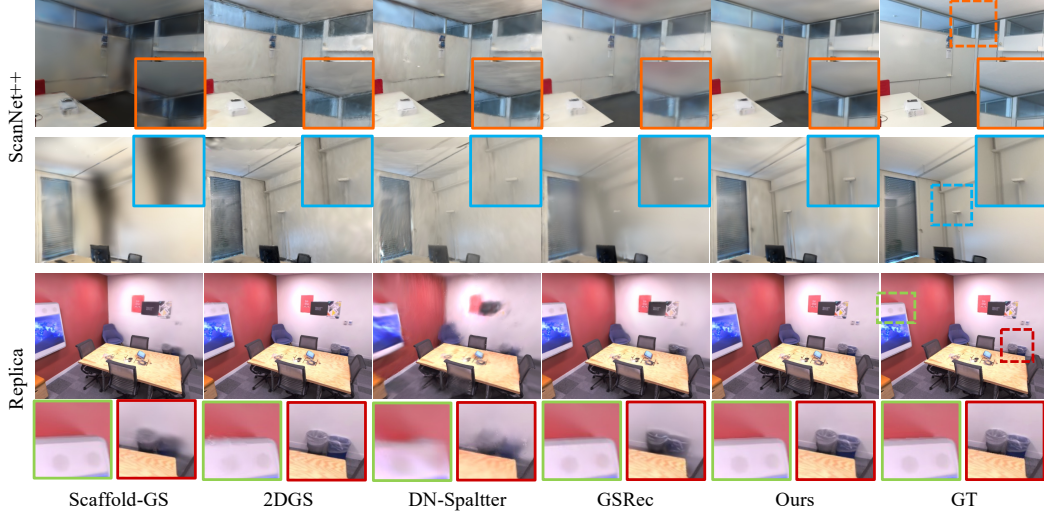


Figure B: **Qualitative comparison of novel view synthesis.** We show the novel view synthesis results of different Gaussian splatting-based approaches on ScanNet++ [13] and Replica [10] datasets. Our method can obtain higher-fidelity rendering results with less noisy information than the baselines.

Table B: **Quantitative comparison of novel view synthesis.** We perform experiments on Replica [10] and ScanNet++ [13] datasets.

Methods	Replica [10]			ScanNet++ [13]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
ScaffoldGS [8]	38.08	0.9660	0.0961	18.25	0.7749	0.2764
2DGS [6]	41.59	0.9823	0.0464	21.87	0.8114	0.3060
DN-Splatter [11]	29.02	0.8967	0.2312	22.76	0.8226	0.2971
GSRec [12]	36.00	0.9574	0.1205	22.96	0.8314	0.2708
Ours	39.58	0.9756	0.0766	22.51	0.8321	0.2517

As described in Sec. 4, we uniformly sample images on the indoor scenes due to redundant images in the original dataset. For each scene in ScanNet [2] and Replica [10], we select one out of every 10 images in the original image sequence. For ScanNet++ [13], we use the image sequence from the iPhone and select one out of every 60 images. All the images are cropped and resized and center cropped to 640×480 . For MatrixCity [7], we use all the provided images and make the image resolution 960×540 . The SfM points are triangulated by COLMAP [9] with given images and corresponding poses.

A.4 Evaluation Metrics

Following previous methods [14, 5], we evaluate accuracy (Acc), completeness (Comp), Chamfer Distance (CD), precision (Prec), recall (Recall), and F1-score on ScanNet [2], ScanNet++ [13], and Replica [10]. Tab. A shows the definition of these metrics.

A.5 Additional Indoor Experiments

Novel View Synthesis. We evaluate the novel view synthesis on Replica [10] and ScanNet++ [13]. Images for training in the experiments are uniformly sampled from the original video sequence. To evaluate the quality of novel view synthesis, we randomly select 50 additional images from the original sequence that are not included in the training set. Tab. B demonstrates the quantitative results, showing that our method produces higher quantitative results than most Gaussian-based methods and reconstructs more accurate surfaces. As illustrated in Fig. B, our method can render photorealistic novel views with accurate geometry, whereas other approaches exhibit artifacts such as Gaussian ellipsoid distortions in the background or blurry objects. While Scaffold-GS and 2DGS perform well on synthetic datasets without significant lighting variations, they struggle to render photorealistic novel views in real scenes. Scaffold-GS models significant lighting variations using

view-dependent geometry, which can lead to overfitting in scenes with substantial lighting changes, such as those in [13]. This overfitting results in inaccurate lighting environment and geometry, as shown in Fig. B. While 2DGS [6] achieves a higher quantitative result on Replica with a discrete representation, the discrete representation exhibits protruding surfaces and results in noisy images on real scenes. GSRec [12] improves geometric accuracy but produces a blurry background and objects, lacking detailed modeling. In contrast, with its precise geometry, our method effectively models lighting variations across views while accurately capturing the appearance of the background and objects.

Semantic Segmentation. We evaluate the semantics from the rendered and the pre-trained segmentation model Mask2Former [1] on Replica [10] and ScanNet++ [13]. As shown in Tab. C, ours achieves better results across all three classes on both datasets. By leveraging Gaussian semantic lifting, our model effectively aggregates multi-view information into 3D space and renders view-consistent semantic maps. In contrast, the 2D semantic segmentation model is more susceptible to image noise, leading to misclassifications, as illustrated in Fig. C. The joint optimization scheme also helps correct semantic misclassifications, particularly around the boundaries between floors and walls.

Table C: **Quantitative comparison of structural layout segmentation on Replica [10] and ScanNet++ [13] dataset.**

Methods	Replica [10]			ScanNet++ [13]		
	IoU _w ↑	IoU _f ↑	IoU _c ↑	IoU _w ↑	IoU _f ↑	IoU _c ↑
Mask2Former [1]	0.628	0.823	0.900	0.684	0.780	0.767
Ours	0.701	0.846	0.927	0.732	0.858	0.777



Figure C: **Qualitative comparison of structural layout segmentation.**

A.6 Additional Qualitative Results

We present qualitative top-view results for ScanNet, ScanNet++, and Replica in Figs. D to F, respectively. For additional comparisons, please refer to our accompanying video.

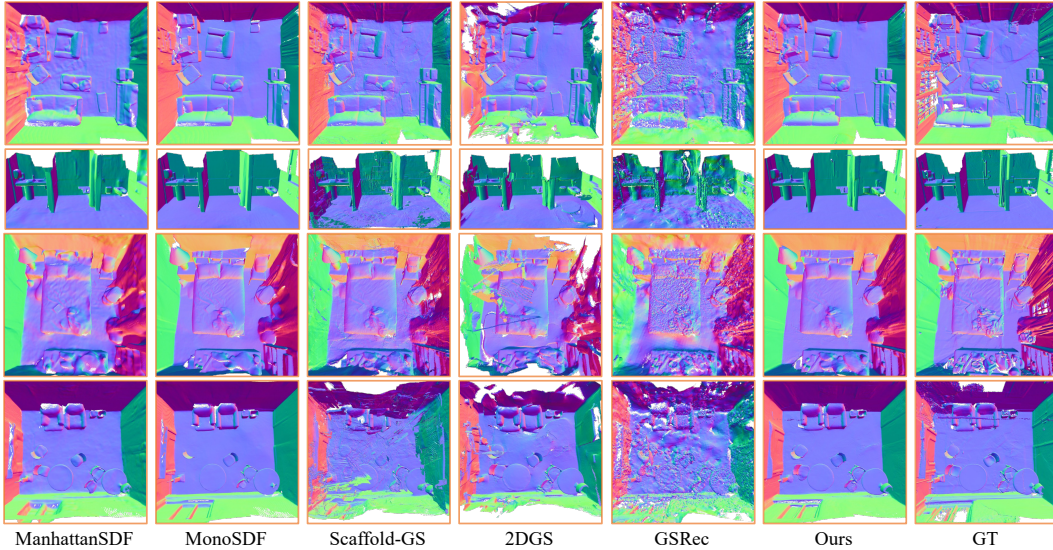


Figure D: **Qualitative comparison of surface reconstruction on ScanNet [2].**

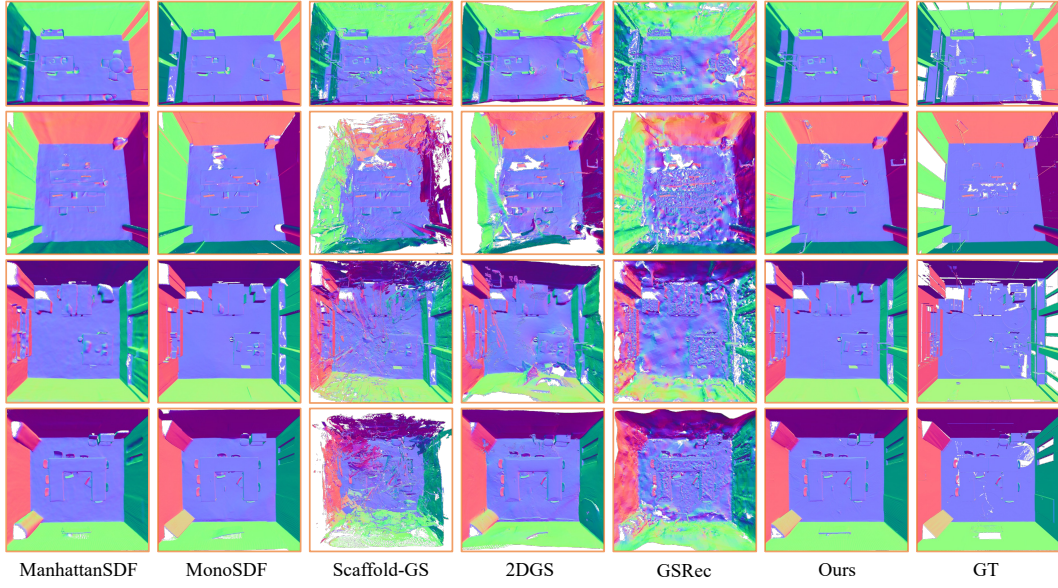


Figure E: **Qualitative comparison of surface reconstruction on ScanNet++ [13].**

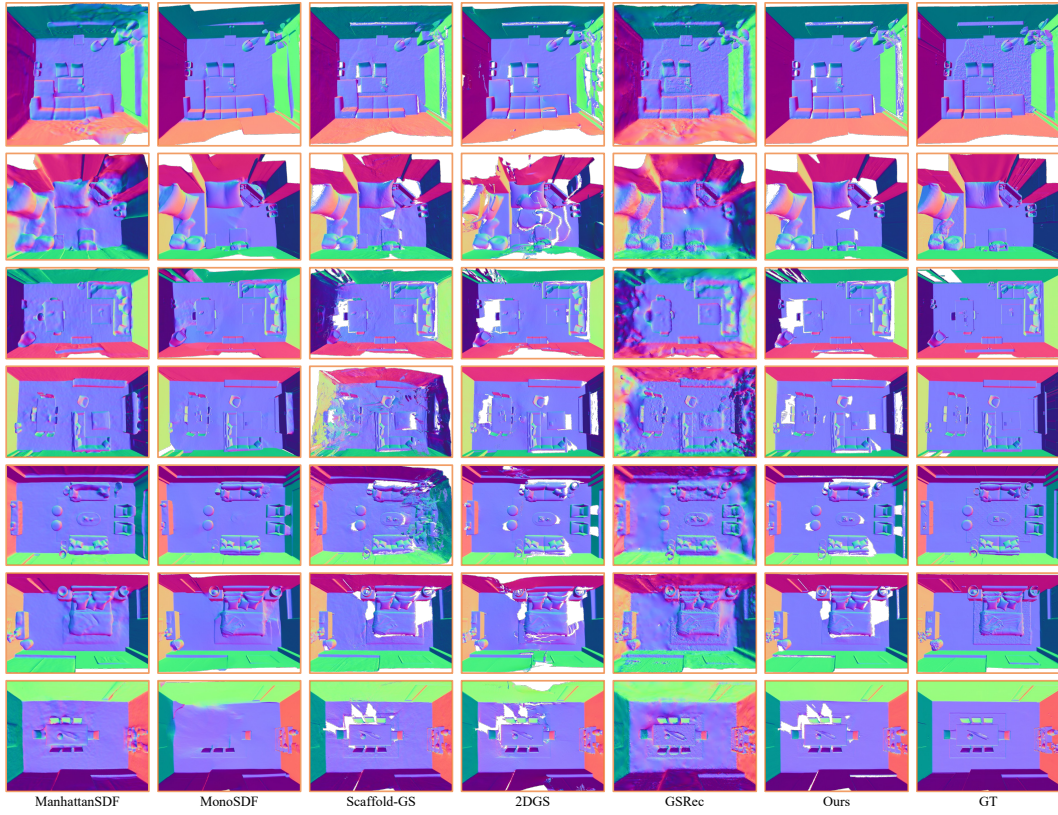


Figure F: **Qualitative comparison of surface reconstruction on Replica [10].**

References

- [1] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1280–1289. IEEE.
- [2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2432–2443, 2017.
- [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 Trans. Graph.*, 36(3):24:1–24:18, 2017.
- [4] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.
- [5] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5511–5520, 2022.
- [6] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024.
- [7] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3205–3215, 2023.
- [8] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20654–20664, 2024.
- [9] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [10] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.
- [11] Matias Turkulainen, Xuqian Ren, Iaroslav Melekhov, Otto Seiskari, Esa Rahtu, and Juho Kannala. Dn-splatter: Depth and normal priors for gaussian splatting and meshing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.
- [12] Qianyi Wu, Jianmin Zheng, and Jianfei Cai. Surface reconstruction from 3d gaussian splatting via local structural hints. In *European Conference on Computer Vision*, pages 441–458. Springer, 2024.
- [13] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023.
- [14] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022.