OmniIndoor3D: Comprehensive Indoor 3D Reconstruction

Xiaobao Wei[1*] **Xiaoan Zhang**[1*] **Hao Wang**[1] **Qingpo Wuwu**[1]
Ming Lu[1] Wenzhao Zheng[2] Shanghang Zhang[1†]
[1]Peking University, Beijing, China [2]University of California, Berkeley, USA
weixiaobao0210@gmail.com

Abstract

Indoor 3D reconstruction is crucial for the navigation of robots within indoor scenes. Current techniques for indoor 3D reconstruction, including truncated signed distance function (TSDF), neural radiance fields (NeRF), and 3D Gaussian Splatting (3DGS), have shown effectiveness in capturing the geometry and appearance of indoor scenes. However, these methods often overlook panoptic reconstruction, which hinders the creation of a comprehensive framework for indoor 3D reconstruction. To this end, we propose a novel framework for comprehensive indoor 3D reconstruction using Gaussian representations, called OmniIndoor3D. This framework enables accurate appearance, geometry, and panoptic reconstruction of diverse indoor scenes captured by a consumer-level RGB-D camera. Since 3DGS is primarily optimized for photorealistic rendering, it lacks the precise geometry critical for high-quality panoptic reconstruction. Therefore, OmniIndoor3D first combines multiple RGB-D images to create a coarse 3D reconstruction, which is then used to initialize the 3D Gaussians and guide the 3DGS training. To decouple the optimization conflict between appearance and geometry, we introduce a lightweight MLP that adjusts the geometric properties of 3D Gaussians. The introduced lightweight MLP serves as a low-pass filter for geometry reconstruction and significantly reduces noise in indoor scenes. To improve the distribution of Gaussian primitives, we propose a densification strategy guided by panoptic priors to encourage smoothness on planar surfaces. Through the joint optimization of appearance, geometry, and panoptic reconstruction, OmniIndoor3D provides comprehensive 3D indoor scene understanding, which facilitates accurate and robust robotic navigation. We perform thorough evaluations across multiple datasets, and OmniIndoor3D achieves state-of-the-art results in appearance, geometry, and panoptic reconstruction. We believe our work bridges a critical gap in indoor 3D reconstruction. The code will be released at: https://ucwxb.github.io/OmniIndoor3D/

1 Introduction

Robotic navigation is a crucial technique in embodied intelligence, necessitating a comprehensive reconstruction of the surrounding environment [1, 41]. The comprehensive reconstruction requires two essential components: spatial perception for obstacle avoidance and localization, and semantic understanding for manipulation and planning [8, 43, 37]. Therefore, a comprehensive 3D reconstruction that captures appearance, geometry, and semantic features is essential for enabling indoor navigation.

[*]Equal contribution
[†]Corresponding author

Preprint. Under review.

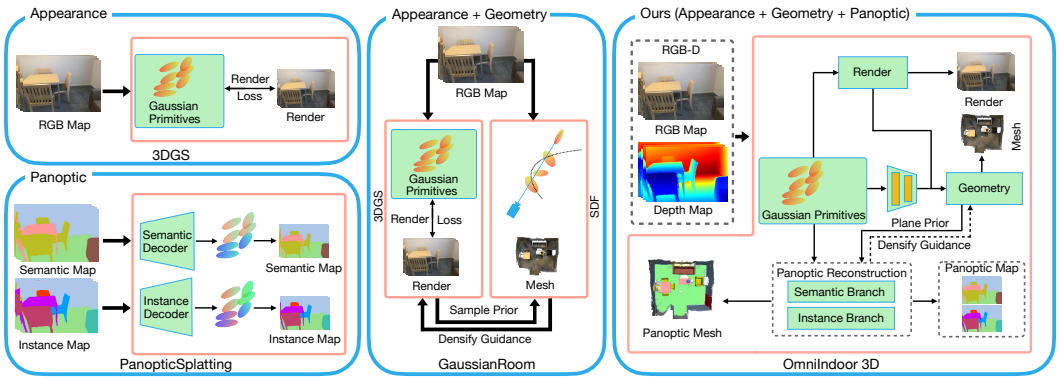

Figure 1: Comparison with existing methods. Unlike previous approaches that treat appearance, geometry, and panoptic reconstruction separately, our OmniIndoor3D presents a unified framework that leverages mutual dependencies for joint optimization, facilitating a comprehensive indoor 3D reconstruction essential for robotic navigation.

Recent advances in 3D reconstruction, particularly Neural Radiance Fields (NeRF) [24] and 3D Gaussian Splatting (3DGS) [15], have led to significant breakthroughs in novel view synthesis and scene representation. Among these, 3DGS stands out because of its outstanding rendering quality and high efficiency [18, 23, 40]. However, the lack of structured geometric constraints in 3DGS often leads to issues such as floating points, hallucinated surfaces, and incomplete reconstructions. Existing methods [35, 12, 7, 44, 53, 49] to address these limitations generally fall into two categories: (1) introducing explicit geometric regularization—e.g., Sugar [12] enforces surface consistency through point-to-surface constraints; and (2) jointly optimizing an auxiliary model, such as a signed distance field (SDF), to guide Gaussian densification, as demonstrated in GaussianRoom [44]. Although these methods explore various strategies for achieving accurate appearance and geometric reconstruction, they overlook the panoptic reconstruction of the scene.

Panoptic scene understanding in 2D has achieved remarkable progress, driven by increasingly powerful network architectures and large-scale annotated datasets [17, 29, 30]. However, extending panoptic segmentation from 2D to 3D remains a challenge. Unlike single-image segmentation, 3D panoptic segmentation [16, 46] requires consistent semantic and instance-level masks across multiple views, which demands a comprehensive understanding of the scene. NeRF-based methods [34, 2, 50, 51] utilize volumetric rendering to achieve promising panoptic segmentation but are limited by high computational cost. 3DGS-based approaches [42, 28, 47, 39, 45] lift 2D segments or features into 3D space, enabling efficient open-vocabulary or panoptic segmentation. However, the precise boundaries for segmentation rely on a clear and accurate geometric reconstruction of the scene. These methods primarily focus on improving rasterized 2D segmentation quality while neglecting the regularization of 3D scene geometry, which ultimately limits their capability for comprehensive 3D scene reconstruction.

To support robotic perception and planning, comprehensive 3D reconstruction is essential. Most existing methods address geometry and panoptic aspects separately, failing to recognize their interdependence. In addition, existing methods fail to exploit the mutual dependencies among appearance, geometry, and panoptic. The conflict between appearance and geometry reconstruction often hinders performance, while accurate geometry and panoptic reconstruction can reinforce each other. A comprehensive framework that simultaneously addresses all three aspects has not yet been thoroughly investigated.

To bridge this gap, we propose OmniIndoor3D, the first framework to enable comprehensive indoor 3D reconstruction based on 3DGS. Given RGB-D images captured by a consumer-level camera, we first perform coarse reconstruction to initialize the Gaussian distribution of OmniIndoor3D. In vanilla 3DGS, jointly optimizing appearance and geometry leads to conflicts, primarily due to the entangled updates of Gaussian scales and rotations. Instead of simultaneously optimizing a time-consuming signed distance field (SDF), we introduce a lightweight multi-layer perceptron (MLP) that learns offsets for scale and rotation parameters. This MLP acts as a low-pass filter, suppressing high-frequency components in 3DGS and producing stable geometric properties, thereby decoupling geometry optimization from appearance refinement. To equip OmniIndoor3D with panoptic reconstruction, we extend each Gaussian with semantic and instance features. These

features are decoded into 3D panoptic labels via a semantic decoder and a set of 3D instance queries. Furthermore, to alleviate blur and noise in RGB-D observations, we propose a panoptic-guided densification strategy that adjusts the Gaussian gradients. The panoptic priors guide the spatial distribution of Gaussians, promoting both completeness and planar smoothness. Through end-to-end optimization, OmniIndoor3D jointly generates high-fidelity appearance, geometry, and panoptic reconstruction (Fig. 1). Navigation robots equipped with standard RGB-D cameras can leverage OmniIndoor3D to achieve comprehensive indoor 3D understanding.

Our contributions are summarized as follows: 1) We present OmniIndoor3D, a novel framework that achieves comprehensive indoor 3D reconstruction using Gaussian representation. 2) To decouple conflicts between geometry and appearance optimization, we propose a lightweight MLP to adjust the geometric properties of 3DGS. 3) To refine the Gaussian distribution and enhance planar smoothness, we introduce a panoptic-guided densification strategy to assist reconstruction with panoptic information. 4) Extensive experiments on ScanNet and ScanNet++ demonstrate that OmniIndoor3D achieves state-of-the-art performance in novel view synthesis, geometric reconstruction, and panoptic lifting.

2 Related Work

Neural Scene Representation. Neural scene representation can be categorized into NeRF-based and 3DGS-based methods. Neural Radiance Fields (NeRF) [24] model scenes as continuous volumetric fields and have shown impressive results in novel view synthesis. Subsequent works improve rendering efficiency [25, 21] and geometric fidelity by incorporating depth supervision, smoothness regularization, and multi-view consistency [10, 26, 36]. However, surfaces extracted from NeRF using Marching Cubes [22] often remain noisy or incomplete. To address this, alternative representations such as occupancy grids [27] and TSDFs [38, 19] have been explored, often guided by SfM priors or geometric constraints [11, 52]. Recently, 3D Gaussian Splatting (3DGS) [15] has emerged as a fast and expressive representation using learnable Gaussian primitives. While efficient for view synthesis, vanilla 3DGS typically relies on SfM initialization [31], leading to suboptimal geometry. To improve spatial distribution, some works utilize LiDAR or RGB-D point clouds [14, 6], while others introduce monocular priors such as depth and normals [44, 53]. Representation extensions include flattened or planar Gaussians [12, 13] and hybrid 2D-3D models [5]. Another direction introduces SDF-constrained optimization [49], jointly refining Gaussians and SDF fields. CarGS [32] further decouples appearance and geometry by identifying scale and rotation as key conflict factors, proposing a geometry-aware MLP. Despite these advancements, most methods remain focused on appearance and geometry reconstruction, lacking the panoptic reconstruction that is critical for robotic perception tasks.

Panoptic Segmentation and Lifting. Panoptic segmentation, first introduced in [16], aims to provide a unified understanding of object instances ("things") and semantic regions ("stuff") in diverse scenes. Early works like UPSNet [46] integrate panoptic segmentation into single networks with novel architectures, extending this concept to 3D has been critical for applications in autonomous driving and robotics. The evolution of 3D panoptic reconstruction has primarily followed two representation paradigms. Neural Radiance Field (NeRF) [24] methods encode scenes into neural networks, offering implicit representations of 3D scene properties including appearance, geometry, and semantics. Several approaches employ NeRFs for 3D panoptic segmentation. Panoptic Lifting [34] proposes a label lifting scheme with linear assignment between predictions and unaligned instance labels. Contrastive Lift [2] achieves 3D object segmentation through contrastive learning. PVLFF [4] builds an instance feature field for open-vocabulary segmentation. PanopticRecon [50, 51] introduces methods for aligning 2D masks and guiding 3D instance segmentation. Despite their promising performance, NeRF-based methods are computationally intensive and thus unsuitable for deployment in real-time robotics applications. Alternatively, 3D Gaussian Splatting (3DGS) [15] methods optimize differentiable Gaussians with real-time rendering speed. Approaches like LEGaussians [33], LangSplat [28], and Feature 3DGS [54] extend Gaussian splatting by adding feature attributes, while Gaussian Grouping [47], OpenGaussian [42], and PLGS [39] focus on instance segmentation through various techniques. PanopticSplatting [45] proposes an end-to-end system for open-vocabulary panoptic reconstruction with query-guided instance segmentation. However, these semantic lifting or feature lifting methods are mainly designed to enhance 2D segmentation, while neglecting the geometric optimization required for accurate 3D semantic mesh reconstruction.

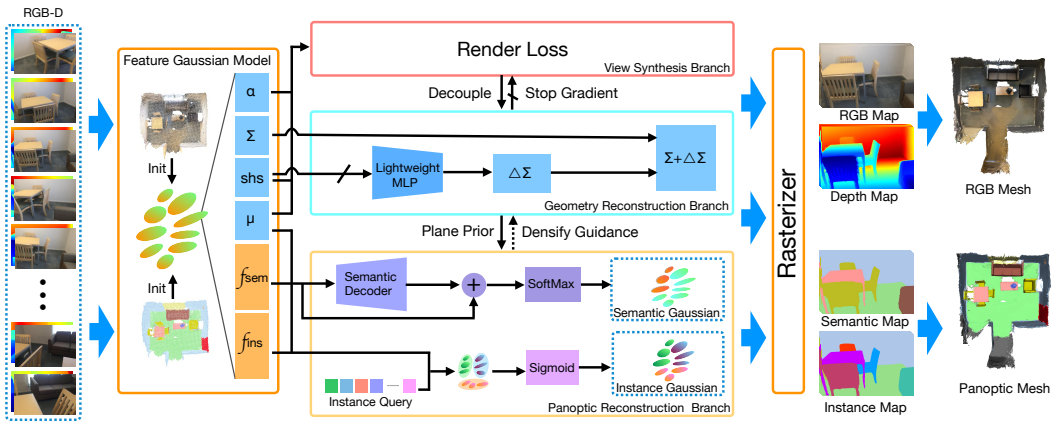

Figure 2: Pipeline of OmniIndoor3D. Given posed RGB-D as inputs, we first extract a coarse 3D reconstruction to initialize the Gaussian distribution. The network subsequently optimizes three dedicated branches, each responsible for novel view synthesis, geometric reconstruction, and panoptic lifting.

We propose OmniIndoor3D, a unified framework that jointly conducts appearance, geometry, and panoptic reconstruction within an efficient Gaussian representation. By decoupling the conflict between appearance and geometry optimization, and leveraging panoptic cues to guide panoptic reconstruction, our method enables high-quality and meaningful indoor 3D reconstruction.

3 Method

3.1 3D Gaussian Initialization and Representation

Gaussian initialization is essential for stable optimization and high-quality rendering. While vanilla 3DGS relies on sparse SfM point clouds, COLMAP is often slow and inaccurate, especially in complex indoor environments. Instead, we utilize RGB-D images captured by consumer-level sensors to reconstruct a coarse but structured point cloud. Multi-view depth maps are projected into the world coordinate frame and aggregated to form a unified point cloud. A voxelization step is applied to reduce noise and control point density. The resulting point cloud is then used to initialize the spatial position μ and the appearance-related spherical harmonics coefficients shs of the Gaussian primitives. To initialize the semantic feature f_{sem} and instance feature f_{ins} of each Gaussian, we follow PanopticSplatting [45] by using Grounded SAM [30] to extract pseudo semantic and instance labels from the input images. These labels are projected into 3D space using the corresponding depth maps, resulting in a 3D point cloud with coarse semantic and instance annotations. We use these labeled points to assign initial semantic and instance features to the Gaussians.

After initialization, the indoor scene is represented as a set of 3D Gaussian primitives. Each Gaussian is defined as a tuple $G = (\Sigma, \mu, shs, \alpha, f_{sem}, f_{ins})$, where $\mu \in \mathbb{R}^3$ is the center, $\Sigma \in \mathbb{R}^{3\times3}$ is the covariance matrix, $shs \in \mathbb{R}^{3(k+1)^2}$ denotes the view-dependent color represented by spherical harmonics of degree $k = 3$, and $\alpha \in \mathbb{R}$ is the opacity. In addition, each Gaussian carries a semantic feature $f_{sem} \in \mathbb{R}^{N_{sem}}$ and an instance feature $f_{ins} \in \mathbb{R}^{N_{ins}}$, where N_{sem} is the number of semantic classes, and N_{ins} is set to a value larger than the maximum number of instances present in the scene.

The spatial density of a Gaussian centered at μ is defined as:

$$G(x) = e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}, \tag{1}$$

where x is a 3D position in the scene. To ensure that the covariance matrix Σ is positive semi-definite, it is constructed as $\Sigma = RSS^\top R^\top$, where $S \in \mathbb{R}^3$ is a diagonal scaling matrix and $R \in \mathbb{R}^{3\times3}$ is a rotation matrix. Each 3D Gaussian is associated with an opacity value α, which modulates its spatial contribution $G(x)$ during the blending process. This weighted contribution is used in both rendering and reconstruction tasks. 3DGS enables efficient scene rendering through tile-based rasterization, avoiding traditional ray marching. Each 3D Gaussian $G(x)$ is projected onto the image plane as a 2D Gaussian, and a tile-based rasterizer composites the scene via α-blending:

$$F(x') = \sum_{i\in N} f_i\sigma_i \prod_{j=1}^{i-1}(1-\sigma_j), \quad \sigma_i = \alpha_i G_i(x), \tag{2}$$

where x' denotes a pixel location in the image plane, and N is the set of 2D Gaussians overlapping that pixel after projection. The value f_i depends on the target task: it is set to the view-dependent color shs_i for appearance rendering, to the depth z_i for geometry reconstruction, and to the semantic or instance feature f_{sem} or f_{ins} for panoptic segmentation.

3.2 Appearance and Geometry Reconstruction

Inspired by prior works [32] identifying the Gaussian covariance as a primary source of conflict between appearance and geometry optimization, we propose to decouple geometry reconstruction via a dedicated lightweight MLP. Specifically, we introduce a geometry-specific covariance adjustment module that acts as a low-pass filter to suppress high-frequency noise in indoor scenes. This module is implemented as a single MLP, which takes as input the detached appearance features and view direction, and predicts a 7D vector representing the scale and rotation (as a quaternion) used to construct the covariance matrix. To avoid mutual interference between the appearance and geometry branches, we denote the input appearance feature as a detached vector ϕ, extracted from the spherical harmonics coefficients shs. Formally, the geometry-adjusted covariance is computed as:

$$\Delta\Sigma = M_{\Sigma}^{geo}(\phi, \theta), \tag{3}$$

where ϕ is the detached appearance feature, θ is the view direction, and M_{Σ}^{geo} is the geometry MLP. The output $\Delta\Sigma \in \mathbb{R}^7$ consists of three scale parameters and four rotation parameters, which are used to construct the geometry-specific covariance matrix for each Gaussian. This formulation enables the geometry branch to refine structural consistency without affecting rendering quality, and proves effective in reducing geometric noise while maintaining surface smoothness. After optimization, we render the RGB image from the optimized Gaussians. The depth map is rendered using the covariance adjusted by the geometry MLP, and is then used to construct a TSDF volume for mesh extraction.

3.3 Panoptic Reconstruction

Semantic Branch. Inspired by prior works [45] on 3D-aware semantic segmentation, we introduce a semantic decoding module that maps each Gaussian's semantic feature f_{sem} into a class probability. To better leverage the pseudo labels assigned during initialization, we design a residual prediction strategy. Specifically, a semantic MLP M_{sem}takes as input semantic features f_{sem} and position μ of the Gaussians, and predicts a correction term to refine the initial semantic logits. Formally, the predicted class logits for one Gaussian are computed as:

$$l_{sem} = \text{softmax}(f_{sem} + M_{sem}(f_{sem}, \mu)), \tag{4}$$

where $l_{sem} \in \mathbb{R}^{N_c}$ is the normalized class probability vector for N_c semantic categories. This residual formulation reduces optimization difficulty and stabilizes learning by preserving the initialization prior. Once the semantic labels of Gaussians are obtained, we perform label blending during the rasterization process. Instead of blending raw features as in [42], we directly blend the predicted class probabilities using the α-blending rule:

$$M_{sem}(x') = \sum_{i \in \mathcal{N}} l_{sem}^{(i)} \, \alpha_i' \prod_{j=1}^{i-1} (1 - \alpha_j'), \tag{5}$$

where x' is a pixel on the image plane, $\mathcal{N}$ is the set of visible Gaussians projected onto x', and α_i' denotes the normalized opacity of the i-th Gaussian. This label blending strategy emphasizes 3D consistency by classifying Gaussians in 3D space before projecting to 2D, which helps mitigate the influence of noisy 2D supervision. Additionally, the softmax normalization prevents Gaussians within the camera frustum from overwhelming the prediction.

Instance Branch. We adopt a query-based design for instance segmentation, inspired by [45], where a set of learnable instance queries interact with scene features to cluster Gaussians into instances. Each instance query is composed of a feature vector $f_q \in \mathbb{R}^d$ and a 3D position $p_q \in \mathbb{R}^3$. The covariance matrix $\Sigma_q \in \mathbb{R}^{3\times3}$ of the instance query is constructed from its scale and rotation parameters. This enables a geometry-aware attention mechanism, where the affinity between a scene Gaussian g and an instance query q is computed by jointly considering feature similarity and spatial proximity.

The attention weight between a scene Gaussian and a query is defined as:

$$A(q, g) = \text{sim}(f_q, f_{ins}^{(g)}) \cdot \mathcal{G}(p_g; p_q, \Sigma_q), \tag{6}$$

where $\text{sim}(f_q, f_{ins}^{(g)})$ denotes feature similarity (e.g., dot product), and $\mathcal{G}(p_g; p_q, \Sigma_q)$ is the probability density of the query's 3D Gaussian evaluated at the scene Gaussian center p_g. Σ_q is constructed from the query's scale and rotation parameters. We apply a softmax over all queries to obtain the instance label distribution for each Gaussian:

$$l_{ins}(g) = \text{softmax}\left(\{A(q_i, g)\}_{i=1}^{N}\right), \tag{7}$$

where N is the total number of instance queries. The final 2D instance map is then rendered via α-blending of these Gaussian-level labels:

$$M_{ins}(x') = \sum_{i \in \mathcal{N}} l_{ins}^{(i)} \alpha_i' \prod_{j=1}^{i-1} (1 - \alpha_j'), \tag{8}$$

where $\mathcal{N}$ denotes the sorted list of Gaussians projected to pixel x'. To improve computational efficiency, we adopt local cross-attention by restricting the query-Gaussian interaction to Gaussians located within the view frustum. Notably, both the semantic and instance rendering processes utilize the geometry-adjusted covariance. The geometry branch provides accurate planar priors that enhance the quality and consistency of segmentation.

3.4 Panoptic-guided Densification

We further improve the spatial distribution of Gaussians through a panoptic-guided densification strategy. Previous geometry-based methods [3] primarily rely on signed distance fields (SDF) to control Gaussian growth. However, these approaches neglect the rich semantic priors available in the scene. To address this, we introduce semantic priors into the densification process. Specifically, we compute a confidence-aware SDF modulation by weighting each Gaussian's SDF value with its semantic confidence. Let $s = d(x) - z(x)$ denote the signed distance function (SDF) value at Gaussian center x, where $d(x)$ is the rendered depth sampled from the image plane, and $z(x)$ is the projected depth of the Gaussian center along the camera ray. The semantic confidence is obtained by taking the softmax over the semantic logits f_{sem} and selecting the maximum non-background class probability. The modulation is defined as:

$$\zeta(s) = \exp\left(-\frac{s^2}{2\sigma^2}\right), \quad \epsilon_g = \nabla_g + \omega_g \cdot \zeta(s) \cdot c_{sem}, \tag{9}$$

where ∇_g is the accumulated gradient magnitude of a Gaussian, c_{sem} is the semantic confidence, and ω_g controls the influence of geometric guidance. A new Gaussian is spawned when ϵ_g exceeds a fixed threshold. This design encourages densification in semantically meaningful and geometrically uncertain regions, improving coverage and segmentation completeness in challenging indoor scenes.

3.5 Training

Appearance Loss. We supervise the rendered image I using a combination of L1 and SSIM losses with respect to the ground truth image I^{gt}:

$$\mathcal{L}_{rgb} = (1 - \lambda_{SSIM})\mathcal{L}_1(I, I^{gt}) + \lambda_{SSIM}\mathcal{L}_{SSIM}(I, I^{gt}), \tag{10}$$

Geometry Loss. To enforce geometric consistency, we supervise the rendered depth $D(x)$ using the ground-truth depth $D^{gt}(x)$ captured by the RGB-D camera:

$$\mathcal{L}_{depth} = \frac{1}{|\mathcal{W}|} \sum_{x \in \mathcal{W}} \|D(x) - D^{gt}(x)\|_1, \tag{11}$$

To further improve global consistency, we introduce a cross-view loss. A pixel x_r in the reference view is projected to a neighboring view via homography H_{rn} and then back-projected using H_{nr}. The consistency is enforced by minimizing the forward-backward reprojection error:

$$\mathcal{L}_{cross} = \frac{1}{|\mathcal{W}|} \sum_{x_r \in \mathcal{W}} \|x_r - H_{nr}H_{rn}x_r\|, \tag{12}$$

Panoptic Loss. For semantic supervision, a cross-entropy loss is applied between the rendered semantic logits M_{sem} and ground truth M_{sem}^{gt}:

$$\mathcal{L}_{sem} = \mathcal{L}_{ce}(M_{sem}, M_{sem}^{gt}), \tag{13}$$

Table 1: Quantitative comparison for novel view synthesis. Results are averaged over the same selected scenes as in GaussianRoom.

Method	ScanNet			ScanNet++		
	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓
3DGS [15]	0.731	22.133	0.387	0.843	21.816	0.294
SuGaR [12]	0.737	22.290	0.382	0.831	20.611	0.318
GaussianPro [7]	0.721	22.676	0.395	0.831	21.285	0.320
DN-Splatter [35]	0.639	21.621	0.312	0.826	20.445	0.268
GaussianRoom [44]	0.758	23.601	0.391	0.844	22.001	0.296
Ours	**0.812**	**25.817**	**0.304**	**0.879**	**25.139**	**0.193**

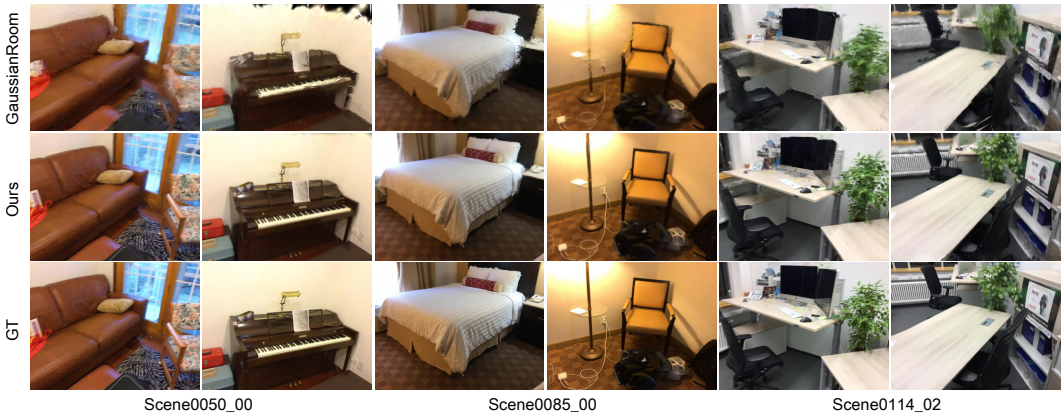

Figure 3: Visualization comparison for novel view synthesis.

For the instance branch, we adopt a combination of Dice loss and binary cross-entropy (BCE) loss between the predicted instance map M_{ins} and the ground truth M_{ins}^{gt}:

$$\mathcal{L}_{ins} = \mathcal{L}_{dice}(M_{ins}, M_{ins}^{gt}) + \mathcal{L}_{bce}(M_{ins}, M_{ins}^{gt}), \tag{14}$$

Total Loss. The total training objective is a weighted sum of all loss components:

$$\mathcal{L}_{total} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{depth}\mathcal{L}_{depth} + \lambda_{cross}\mathcal{L}_{cross} + \lambda_{sem}\mathcal{L}_{sem} + \lambda_{ins}\mathcal{L}_{ins}, \tag{15}$$

In our experiments, the weights are empirically set as $\lambda_{rgb} = 1.0$, $\lambda_{depth} = 1.0$, $\lambda_{cross} = 1.5$, $\lambda_{sem} = 0.5$, and $\lambda_{ins} = 0.5$.

4 Experiments

4.1 Experimental Setup

Datasets. We validate our approach on two publicly available indoor scene datasets, ScanNet [9] and ScanNet++ [48]. To assess reconstruction and rendering quality, and compare with current state-of-the-art methods, we select the same scenes as GaussianRoom [44], a total of 10 indoor scenes, 8 scenes from ScanNet, and 2 scenes from ScanNet++. To evaluate the panoptic segmentation performance, we follow PanopticSplatting [45] and select 7 indoor scenes, including 4 from ScanNet and 3 from ScanNet++. We strictly follow the experimental settings used in the baseline methods, including the training and validation splits as well as the evaluation tools. More details on dataset preprocessing are provided in the appendix.

Evaluation metrics. We adopt standard evaluation metrics for each task. For novel view synthesis, we report SSIM, PSNR, and LPIPS to measure image quality. For geometric reconstruction, we use Accuracy (Acc.), Completion (Com.), Precision (Pre.), Recall (Re.), and F1-score (F1). For panoptic lifting, we evaluate with PQ, SQ, RQ, mIoU, mAcc, mCov, and mW-Cov. Due to space limitations, detailed definitions of all metrics are provided in the appendix.

Table 2: Quantitative comparison for geometric reconstruction. Results are averaged over the same selected scenes as in GaussianRoom.

Method	ScanNet					ScanNet++				
	Acc.↓	Com.↓	Pre.↑	Re.↑	F1↑	Acc.↓	Com.↓	Pre.↑	Re.↑	F1↑
COLMAP [31]	0.062	0.090	0.640	0.569	0.600	0.091	0.093	0.519	0.520	0.517
NeRF [24]	0.160	0.065	0.378	0.576	0.454	0.135	0.082	0.421	0.569	0.484
NeuS [38]	0.105	0.124	0.448	0.378	0.409	0.163	0.196	0.316	0.265	0.288
MonoSDF [52]	0.048	0.068	0.673	0.558	0.609	0.039	0.043	0.816	0.840	0.827
HelixSurf [20]	0.063	0.134	0.657	0.504	0.567	——	——	——	——	——
3DGS [15]	0.338	0.406	0.129	0.067	0.085	0.113	0.790	0.445	0.103	0.163
GaussianPro [7]	0.313	0.394	0.112	0.075	0.088	0.141	1.283	0.353	0.081	0.129
SuGaR [12]	0.167	0.148	0.361	0.373	0.366	0.129	0.121	0.435	0.444	0.439
DN-Splatter [35]	0.212	0.210	0.153	0.182	0.166	0.294	0.276	0.108	0.108	0.107
2DGS [13]	0.167	0.152	0.311	0.341	0.324	——	——	——	——	——
GaussianRoom [44]	0.047	0.043	0.800	0.739	0.768	0.035	0.037	0.894	0.852	0.872
Ours	**0.023**	**0.024**	**0.927**	**0.907**	**0.917**	**0.008**	**0.011**	**0.996**	**0.970**	**0.983**

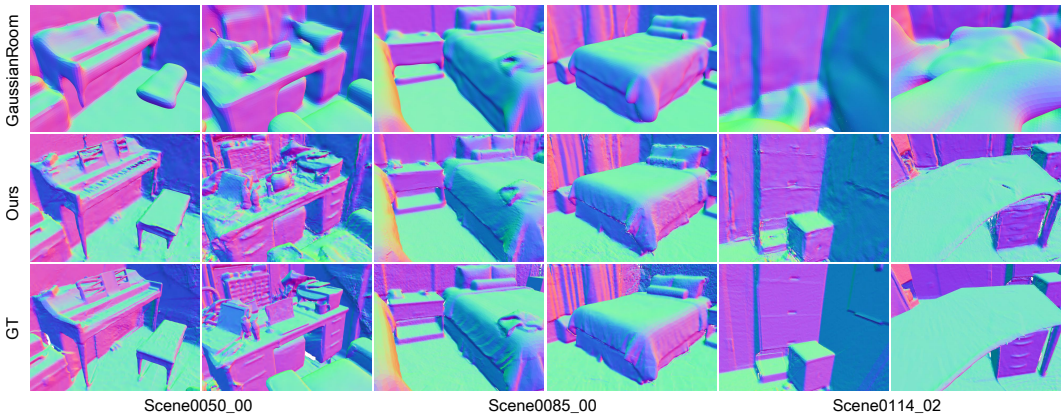

Figure 4: Visualization comparison for geometric reconstruction.

4.2 Novel View Synthesis

We evaluate OmniIndoor3D on novel view synthesis against leading 3DGS-based surface reconstruction methods. As shown in Tab. 1, our approach consistently outperforms baselines across metrics on both ScanNet and ScanNet++. GaussianRoom [44] suffers from low performance, as its SDF-guided pruning removes detail-preserving Gaussians and causes over-smoothing. Our improved performance stems from two key components: (1) RGB-D fusion for Gaussian initialization provides structured geometric priors, reducing noise and ambiguity during early optimization; (2) a lightweight MLP, guided by depth regularization, adjusts Gaussian geometry and effectively decouples appearance from geometry optimization. This prevents high-frequency noise and preserves visual fidelity. As shown in Fig. 3, OmniIndoor3D generates sharper textures and more consistent multi-view structures, while GaussianRoom often suffers from rendering holes due to over-pruning.

4.3 Geometric Reconstruction

As shown in Tab. 2, our method outperforms all baselines and achieves state-of-the-art results in all evaluated scenes. OmniIndoor3D significantly improves metrics across both datasets, indicating that our approach preserves more accurate surface details while maintaining structural completeness. These improvements can be attributed to the use of RGB-D fusion for Gaussian initialization and our depth-guided geometric regularization. Additionally, we adopt a panoptic-guided densification strategy to densify Gaussians near planar regions. This encourages Gaussians to be cloned and split at appropriate locations. From the visual comparisons in Fig. 4, we observe that our method reconstructs clearer object boundaries and finer surface details. Mesh reconstructed by OmniIndoor3D exhibits fewer noisy fluctuations and better preservation of flat and planar structures. In contrast, GaussianRoom suffers from over-smoothed surfaces and distorted geometry, particularly in thin or high-frequency regions. This degradation stems from its reliance on an SDF field for mesh extraction. Moreover, our framework remains efficient. Instead of relying on an additional SDF field with high

Table 3: Quantitative comparison for panoptic lifting. Results are averaged over the same selected scenes as in PanopticSplatting.

Method	ScanNet							ScanNet++						
	PQ↑	SQ↑	RQ↑	mIoU↑	mAcc↑	mCov↑	mW-Cov↑	PQ↑	SQ↑	RQ↑	mIoU↑	mAcc↑	mCov↑	mW-Cov↑
Panoptic Lifting [34]	57.86	61.96	85.31	67.91	78.59	45.88	59.93	71.14	77.48	88.14	81.34	89.67	56.17	68.51
Contrastive Lift [2]	37.35	41.91	57.60	64.77	75.80	13.21	23.26	47.58	57.23	65.81	81.09	89.30	27.39	36.51
PVLFF [4]	30.11	51.71	44.43	55.41	63.96	45.75	48.41	52.24	66.86	65.56	62.53	70.31	67.95	75.47
PanopticRecon [50]	63.70	64.81	81.17	68.62	80.87	66.58	77.84	68.29	77.01	85.05	77.75	87.08	51.34	62.79
Gaussian Grouping [47]	43.75	50.63	72.68	58.05	68.68	52.70	58.10	33.10	40.60	67.27	59.53	68.13	29.83	36.83
OpenGaussian [42]	48.73	51.48	88.10	54.05	68.43	44.43	49.60	51.03	56.93	85.73	61.80	73.97	50.00	51.02
PanopticSplatting [45]	74.75	74.75	**100.0**	74.95	83.70	73.18	79.63	77.73	82.70	93.60	81.90	89.50	74.73	78.03

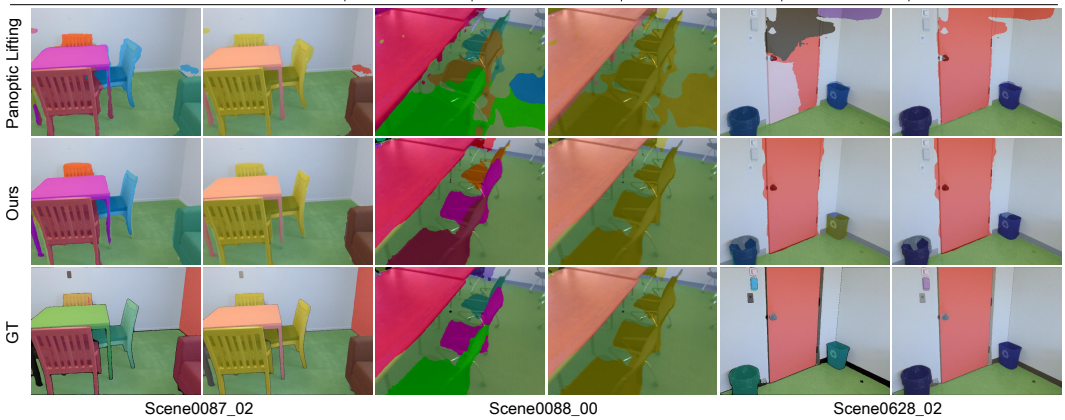

Figure 5: Visualization comparison for panoptic lifting.

memory and computational cost, we employ a lightweight MLP that reduces the interference of geometry optimization on appearance rendering.

4.4 Panoptic Lifting

We evaluate the panoptic lifting performance of our method on the ScanNet and ScanNet++ datasets. As shown in Tab. 3, our approach outperforms all baselines across different evaluation metrics. Compared to Panoptic Lifting [34] and Contrastive Lift [2], our method achieves significantly better segmentation quality. These NeRF-based methods suffer from slow training and rendering time. 3DGS-based methods such as OpenGaussian [42] and PanopticSplatting [45] improve efficiency but still struggle with accurate geometry. These models mainly lift 2D predictions into 3D, leading to inconsistent instance boundaries and fragmented labels. As illustrated in Fig. 5, our method delivers sharper instance boundaries and more complete segmentations. This improvement comes from our RGB-D fusion for Gaussian initialization and panoptic-guided densification strategy. By incorporating semantic priors during optimization, we regularize the Gaussian distribution and promote planar consistency. Overall, OmniIndoor3D provides a unified solution for appearance, geometry, and panoptic reconstruction.

Table 4: Component-wise ablation study. Evaluated on the same scenes as in Tab. 3.

Method	Novel View Synthesis			Geometric Reconstruction					Panoptic Segmentation						
	SSIM↑	PSNR↑	LPIPS↓	Acc.↓	Com.↓	Pre.↑	Re.↑	F1↑	PQ↑	SQ↑	RQ↑	mIoU↑	mAcc↑	mCov↑	mW-Cov↑
w/o RGB-D Init	0.844	25.728	0.315	0.102	0.039	0.581	0.796	0.663	37.25	45.27	52.77	68.95	72.48	50.31	58.05
w/o Geo Decouple	0.874	29.899	0.263	0.021	0.019	0.936	0.938	0.937	74.68	88.55	80.84	77.72	84.77	68.46	77.92
w/o Pan-guided	0.858	30.356	0.285	0.024	**0.017**	0.934	**0.949**	0.941	80.55	87.74	87.74	**77.80**	78.82	67.49	76.60
Full Model	**0.890**	**30.956**	**0.250**	**0.019**	0.018	**0.949**	0.948	**0.948**	**84.60**	**88.74**	**94.69**	76.19	**87.42**	**81.08**	**83.96**

4.5 Component-wise Ablation Study

To assess the contribution of each key component in OmniIndoor3D, we conduct an ablation study across the three core tasks: novel view synthesis, geometric reconstruction, and panoptic lifting on the 4 ScanNet scenes. The results are summarized in Tab. 4. Removing RGB-D initialization leads to a disorganized Gaussian distribution and reduced reconstruction quality, as the lack of structured depth priors introduces ambiguity in early optimization.

Without the proposed MLP-based geometry decoupling module, appearance and geometry optimization interfere with each other, causing degradation across all tasks. Although geometry remains

relatively accurate, the rendering performance decreases. Panoptic segmentation also suffers from the conflict between them, highlighting the importance of separating the two objectives.

Disabling panoptic-guided densification primarily affects panoptic segmentation by reducing semantic coverage and consistency, showing that semantic priors are essential for guiding Gaussian growth toward meaningful regions. Additional ablation studies are provided in the appendix.

5 Conclusion

We introduce OmniIndoor3D, the first unified framework for comprehensive indoor 3D reconstruction, which simultaneously performs appearance, geometry, and panoptic reconstruction using Gaussian representations. Extensive experiments across multiple benchmarks demonstrate that OmniIndoor3D achieves state-of-the-art performance in novel view synthesis, geometric reconstruction, and panoptic lifting. Our approach enables robust and consistent 3D scene understanding, facilitating improved robotic navigation within complex indoor environments.

References

[1] Raghad Alqobali, Maha Alshmrani, Reem Alnasser, Asrar Rashidi, Tareq Alhmiedat, and Osama Moh'd Alia. A survey on robot semantic navigation systems for indoor environments. *Applied Sciences*, 14(1):89, 2023.

[2] Yash Bhalgat, Iro Laina, João F Henriques, Andrea Vedaldi, and Andrew Zisserman. Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion. In *NeurIPS*, 2023.

[3] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[4] Haoran Chen, Kenneth Blomqvist, Francesco Milano, and Roland Siegwart. Panoptic vision-language feature fields. *IEEE Robotics and Automation Letters*, 9(3):2144–2151, 2024.

[5] Peng Chen, Xiaobao Wei, Qingpo Wuwu, Xinyi Wang, Xingyu Xiao, and Ming Lu. Mixedgaussianavatar: Realistically and geometrically accurate head avatar via mixed 2d-3d gaussian splatting. *arXiv preprint arXiv:2412.04955*, 2024.

[6] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, et al. Omnire: Omni urban scene reconstruction. *arXiv preprint arXiv:2408.16760*, 2024.

[7] Kai Cheng, Xiaoxiao Long, Kaizhi Yang, Yao Yao, Wei Yin, Yuexin Ma, Wenping Wang, and Xuejin Chen. Gaussianpro: 3d gaussian splatting with progressive propagation. In *Forty-first International Conference on Machine Learning*, 2024.

[8] Jonathan Crespo, Jose Carlos Castillo, Oscar Martinez Mozos, and Ramon Barber. Semantic information for robot navigation: A survey. *Applied Sciences*, 10(2):497, 2020.

[9] 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.

[10] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12882–12891, 2022.

[11] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 35:3403–3416, 2022.

[12] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5354–5363, 2024.

[13] 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.

[14] Nan Huang, Xiaobao Wei, Wenzhao Zheng, Pengju An, Ming Lu, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. S3gaussian: Self-supervised street gaussians for autonomous driving. *arXiv preprint arXiv:2405.20323*, 2024.

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[16] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019.

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[18] Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. Compact 3d gaussian representation for radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21719–21728, 2024.

[19] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023.

[20] Zhihao Liang, Zhangjin Huang, Changxing Ding, and Kui Jia. Helixsurf: A robust and efficient neural implicit surface learning of indoor scenes with iterative intertwined regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13165–13174, 2023.

[21] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020.

[22] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998.

[23] 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.

[24] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[25] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.

[26] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5480–5490, 2022.

[27] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3504–3515, 2020.

[28] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024.

[29] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[30] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[31] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016.

[32] You Shen, Zhipeng Zhang, Xinyang Li, Yansong Qu, Yu Lin, Shengchuan Zhang, and Liujuan Cao. Evolving high-quality rendering and reconstruction in a unified framework with contribution-adaptive regularization. *arXiv preprint arXiv:2503.00881*, 2025.

[33] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5333–5343, 2024.

[34] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9043–9052, 2023.

[35] 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 *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2421–2431. IEEE, 2025.

[36] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9065–9076, 2023.

[37] Hao Wang, Xiaobao Wei, Xiaoan Zhang, Jianing Li, Chengyu Bai, Ying Li, Ming Lu, Wenzhao Zheng, and Shanghang Zhang. Embodiedocc++: Boosting embodied 3d occupancy prediction with plane regularization and uncertainty sampler. *arXiv preprint arXiv:2504.09540*, 2025.

[38] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 27171–27183, 2021.

[39] Yu Wang, Xiaobao Wei, Ming Lu, and Guoliang Kang. Plgs: Robust panoptic lifting with 3d gaussian splatting. *arXiv preprint arXiv:2410.17505*, 2024.

[40] Xiaobao Wei, Peng Chen, Guangyu Li, Ming Lu, Hui Chen, and Feng Tian. Gazegaussian: High-fidelity gaze redirection with 3d gaussian splatting. *arXiv preprint arXiv:2411.12981*, 2024.

[41] Liyana Wijayathunga, Alexander Rassau, and Douglas Chai. Challenges and solutions for autonomous ground robot scene understanding and navigation in unstructured outdoor environments: A review. *Applied Sciences*, 13(17):9877, 2023.

[42] Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, et al. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. *arXiv preprint arXiv:2406.02058*, 2024.

[43] Yuqi Wu, Wenzhao Zheng, Sicheng Zuo, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Embodiedocc: Embodied 3d occupancy prediction for vision-based online scene understanding. *arXiv preprint arXiv:2412.04380*, 2024.

[44] Haodong Xiang, Xinghui Li, Kai Cheng, Xiansong Lai, Wanting Zhang, Zhichao Liao, Long Zeng, and Xueping Liu. Gaussianroom: Improving 3d gaussian splatting with sdf guidance and monocular cues for indoor scene reconstruction. *arXiv preprint arXiv:2405.19671*, 2024.

[45] Yuxuan Xie, Xuan Yu, Changjian Jiang, Sitong Mao, Shunbo Zhou, Rui Fan, Rong Xiong, and Yue Wang. Panopticsplatting: End-to-end panoptic gaussian splatting. *arXiv preprint arXiv:2503.18073*, 2025.

[46] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8818–8826, 2019.

[47] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European Conference on Computer Vision*, pages 162–179. Springer, 2024.

[48] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023.

[49] Mulin Yu, Tao Lu, Linning Xu, Lihan Jiang, Yuanbo Xiangli, and Bo Dai. Gsdf: 3dgs meets sdf for improved rendering and reconstruction. *arXiv preprint arXiv:2403.16964*, 2024.

[50] Xuan Yu, Yili Liu, Chenrui Han, Sitong Mao, Shunbo Zhou, Rong Xiong, Yiyi Liao, and Yue Wang. Panopticrecon: Leverage open-vocabulary instance segmentation for zero-shot panoptic reconstruction. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12947–12954. IEEE, 2024.

[51] Xuan Yu, Yuxuan Xie, Yili Liu, Haojian Lu, Rong Xiong, Yiyi Liao, and Yue Wang. Leverage cross-attention for end-to-end open-vocabulary panoptic reconstruction. *arXiv preprint arXiv:2501.01119*, 2025.

[52] 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.

[53] Wanting Zhang, Haodong Xiang, Zhichao Liao, Xiansong Lai, Xinghui Li, and Long Zeng. 2dgs-room: Seed-guided 2d gaussian splatting with geometric constrains for high-fidelity indoor scene reconstruction. *arXiv preprint arXiv:2412.03428*, 2024.

[54] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024.