2DGS-Room: Seed-Guided 2D Gaussian Splatting with Geometric Constraints for High-Fidelity Indoor Scene Reconstruction

1st Wanting Zhang
Shenzhen International Graduate School
Tsinghua University
Shenzhen, China
zhangwt23@mails.tsinghua.edu.cn

2nd Haodong Xiang
Shenzhen International Graduate School
Tsinghua University
Shenzhen, China
xhd22@tsinghua.org.cn

3rd Zhichao Liao
Shenzhen International Graduate School
Tsinghua University
Shenzhen, China
liaozc23@mails.tsinghua.edu.cn

4th Xiansong Lai
Shenzhen International Graduate School
Tsinghua University
Shenzhen, China
laixs21@tsinghua.org.cn

5th Xinghui Li*
Shenzhen International Graduate School
Tsinghua University
Shenzhen, China
li-xh21@tsinghua.org.cn

6th Long Zeng*
Shenzhen International Graduate School
Tsinghua University
Shenzhen, China
zenglong@sz.tsinghua.edu.cn

***Abstract*—The reconstruction of indoor scenes is a critical task for embodied robotics, where accurate perception of geometry is essential for navigation, scene understanding, and spatial reasoning. Although recent 3D Gaussian Splatting (3DGS) methods show promising results in novel view synthesis, their performance degrades in textureless regions and complex indoor layouts, resulting in incomplete and noisy reconstructions. In this paper, we introduce 2DGS-Room, a novel method leveraging 2D Gaussian Splatting for high-fidelity indoor scene reconstruction. Specifically, we introduce a seed-guided strategy to control the distribution of 2D Gaussians through dynamic growth and pruning, while depth and normal priors enhance geometry in both detailed and low-texture regions. Furthermore, multi-view consistency constraints improve structural robustness across views. Extensive experiments demonstrate that our method achieves state-of-the-art performance in indoor scene reconstruction. Project page: *https://valentina-zhang.github.io/2DGS-Room/*.**

***Index Terms*—Indoor Scene Reconstruction, 2D Gaussian Splatting, Embodied Robotics.**

I. Introduction

3D reconstruction from multi-view RGB images is a fundamental task in computer vision and computer graphics, with applications in areas such as virtual reality, autonomous driving, and robotics. It plays a crucial role in bridging the gap between simulated environments and real-world scenarios [1], [2]. However, reconstructing indoor scenes presents unique challenges due to the complexity of these environments, which often contain large textureless regions.

MVS-based methods [3], [4] often produce incomplete or distorted reconstructions in indoor scenes due to geometric ambiguities caused by textureless regions. Neural radiance field approaches [5]–[8] mitigate this issue by modeling scenes with signed distance fields (SDFs), enabling accurate and complete mesh recovery through the continuity of neural SDFs and the integration of monocular geometric priors [6]. However, their reliance on dense ray sampling leads to high computational costs and long optimization times. 3D Gaussian Splatting (3DGS) [9] improves rendering and optimization efficiency via differentiable rasterization, opening new opportunities for efficient reconstruction. Building on this idea, 2DGS [10] employs planar 2D Gaussians to better capture scene surfaces, significantly enhancing reconstruction quality. Despite these advances, existing Gaussian-splatting-based methods still suffer from floating artifacts and missing structures in indoor environments, primarily due to the absence of structured geometric constraints.

In this work, we propose 2DGS-Room, a novel framework for high-fidelity indoor scene reconstruction based on 2D Gaussian Splatting. Considering the scene's underlying structure, we design a seed-guided mechanism to control the distribution and density of 2D Gaussians. It consists of a seed-guided initialization that aligns Gaussians with scene surfaces for improved geometric accuracy, and a seed-guided optimization strategy that dynamically adjusts seed density via gradient-guided growth and contribution-based pruning, enabling efficient modeling of fine details. We further incorporate monocular depth and normal priors as geometric constraints. The depth prior corrects distortions in detailed regions, while the normal prior enhances surface estimation in textureless areas. Finally, we introduce multi-view geometric and photometric consistency constraints to suppress residual artifacts across views. Extensive qualitative and quantitative experiments show that compared with Gaussian-based methods, 2DGS-Room achieves start-of-the-art performance in indoor scenarios.

*Corresponding Author.

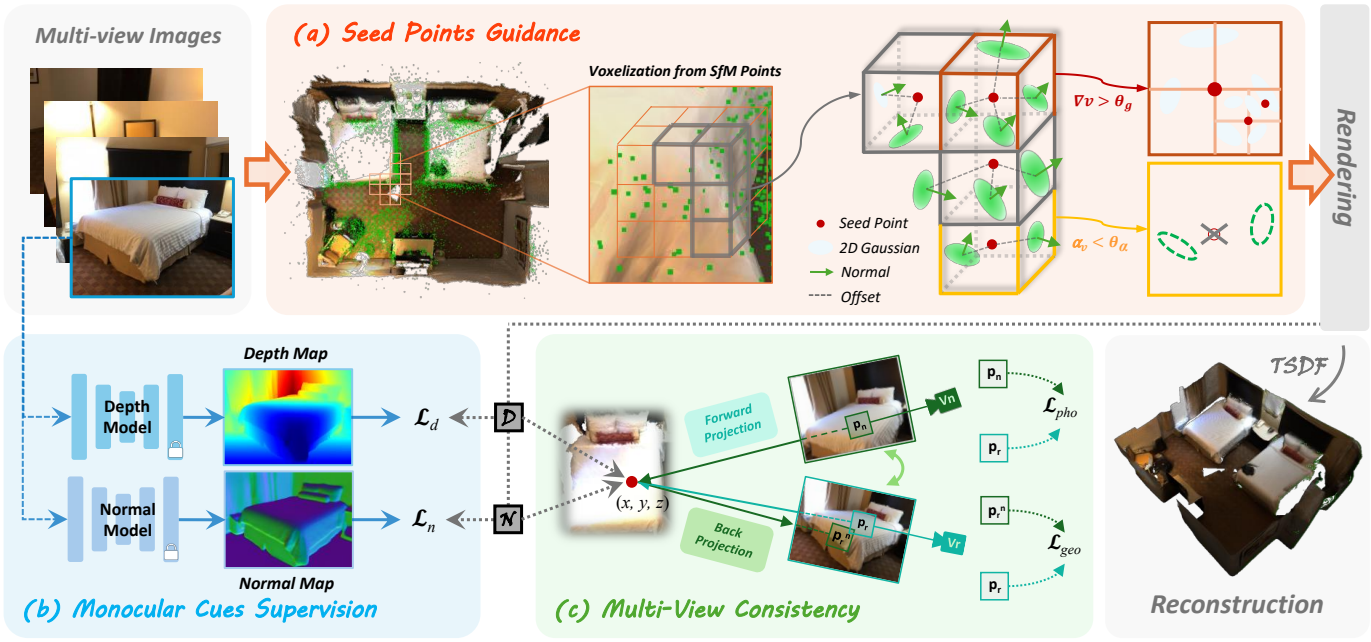

Fig. 1: Overview of 2DGS-Room. Given multi-view posed images, we improve 2DGS to achieve high-fidelity geometric reconstruction for indoor scenes.

In summary, our contributions are as follows:

- We propose **2DGS-Room**, a novel method for indoor scene reconstruction based on 2DGS, which leverages the seed points maintaining the scene structure to guide the distribution and density of 2D Gaussians.
- We introduce monocular depth and normal priors to provide geometric cues, improving the reconstruction of detailed areas and textureless regions, respectively.
- We employ multi-view constraints incorporating geometric and photometric consistency to further enhance the reconstruction quality.
- Our method achieves high-quality surface reconstruction for indoor scenes. Extensive experiments on indoor scene datasets show that our method achieves state-of-the-art in multiple evaluation metrics.

II. Related Work

A. Multi-View Stereo

Multi-view stereo (MVS) methods [3] reconstruct 3D scenes by estimating pixel depths via feature matching across posed images, followed by Poisson surface reconstruction. In indoor scenes, particularly in large texture-less regions, they struggle due to insufficient visual features. Voxel-based approaches [11] avoid feature matching by optimizing occupancy and color in a voxel grid, but suffer from memory limitations at high resolutions. Learning-based MVS methods [4] implicitly match multi-view features with neural networks for end-to-end reconstruction, yet still fail under occlusions, complex lighting, or low-texture conditions despite large-scale training.

B. Neural Radiance Field

Neural Radiance Fields (NeRF) [12] employs a multi-layer perceptron (MLP) to model a continuous volumetric function of density and color, enabling novel view synthesis through volume rendering. Methods such as Mip-NeRF [13] enhance rendering quality by improving the ray sampling strategy. Depth regularization [14] and smoothness constraints [15] improve rendering efficiency and quality, while multi-view consistency regularization aids sparse-view synthesis [16]. Alternative implicit functions, including occupancy grids [17] and signed distance functions (SDFs) [18], [19], replace NeRF's density field for better geometric reconstruction. To further enhance reconstruction quality, [20] suggests regularizing optimization with SfM points, while [6], [21] incorporate priors like the Manhattan world assumption and pseudo depth supervision. However, these approaches often lead to incomplete reconstructions and require extensive optimization time.

C. Gaussian Splatting

3D Gaussian Splatting (3DGS) [9] represents 3D scenes with learnable Gaussian primitives, enabling high-quality novel view synthesis with fast training and rendering. Initialized from sparse SfM point clouds [22], it optimizes only with image loss, leading to unstructured Gaussian distributions and poor geometric fidelity. Subsequent works such as DN-Splatter [23], GaussianRoom [24], and GSDF [25] incorporate geometric priors or SDF supervision to improve geometry. SuGaR [26], PGSR [27], and RaDe-GS [28] adopt flattened Gaussians to enhance surface reconstruction. 2DGS [10] further simplifies the representation by using planar Gaussians, achieving improved geometry, but still suffers in indoor scenes

due to the lack of geometric constraints. Scaffold-GS [29] introduces structural guidance for better rendering, but remains limited in geometric accuracy.

III. Preliminary

2DGS [10] models scenes using 2D elliptical Gaussians (surfels) instead of 3D volumetric ones. Each 2D Gaussian is defined in a local tangent plane by a center $\mathbf{p}_k$, two orthogonal tangential vectors $\mathbf{t}_u$, $\mathbf{t}_v$, and scaling factors (s_u, s_v). The normal is $\mathbf{t}_w = \mathbf{t}_u \times \mathbf{t}_v$, forming a rotation matrix $\mathbf{R} = [\mathbf{t}_u, \mathbf{t}_v, \mathbf{t}w]$ and a scaling matrix $\mathbf{S} = \mathrm{diag}(s_u, s_v, 0)$. Then a 2D Gaussian can be parameterized:

$$P(u,v) = \mathbf{p}k + s_u\mathbf{t}_u u + s_v\mathbf{t}_v v = \mathbf{H}(u,v,1,1), \tag{1}$$

with the homogeneous transformation matrix:

$$\mathbf{H} = \begin{bmatrix} s_u\mathbf{t}_u & s_v\mathbf{t}_v & \mathbf{0} & \mathbf{p}_k \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{RS} & \mathbf{p}_k \\ \mathbf{0} & 1 \end{bmatrix}. \tag{2}$$

In the local (u, v) space, the Gaussian function is:

$$\mathcal{G}(\mathbf{u}) = \exp\left(-\frac{u^2+v^2}{2}\right). \tag{3}$$

For efficient rendering, each 2D Gaussian is projected onto the image plane using the world-to-screen matrix $\mathbf{W}$:

$$\mathbf{x} = (xy, yz, z, z)^\top = \mathbf{WH}(u,v,1,1)^\top, \tag{4}$$

where $\mathbf{x}$ is the projected point in screen space.

To improve numerical stability, ray-splat intersections are computed via three-plane intersections. Given an image coordinate $\mathbf{x} = (x, y)$, the ray of a pixel can be defined by the intersection of two homogeneous planes:

$$\mathbf{h}_x = (-1, 0, 0, x), \quad \mathbf{h}_y = (0, -1, 0, y), \tag{5}$$

which are transformed to the Gaussian's tangent frame:

$$\mathbf{h}_u = (\mathbf{WH})^\top \mathbf{h}_x, \quad \mathbf{h}_v = (\mathbf{WH})^\top \mathbf{h}_y. \tag{6}$$

The intersection point $(u(\mathbf{x}), v(\mathbf{x}))$ is then computed as:

$$u(\mathbf{x}) = \frac{h_u^2 h_v^4 - h_u^4 h_v^2}{h_u^1 h_v^2 - h_u^2 h_v^1}, \quad v(\mathbf{x}) = \frac{h_u^4 h_v^1 - h_u^1 h_v^4}{h_u^1 h_v^2 - h_u^2 h_v^1}. \tag{7}$$

IV. Method

Given multi-view posed images, our goal is to optimize 2DGS [10] to accurately reconstruct the geometry of indoor scenes. To this end, we first propose a seed-guided mechanism that uses seed points to control the distribution and density of 2D Gaussians, improving both accuracy and efficiency (Sec. IV-A). To further enhance geometric fidelity, we incorporate depth and normal priors to better represent detailed and textureless regions, respectively (Sec. IV-B). Finally, to reduce floating artifacts from lighting variations, we introduce multi-view consistency constraints (Sec. IV-C). An overview of the framework is shown in Fig. 1.

A. Seed Points Guidance

Existing methods [9], [10] tend to optimize Gaussians relying on each training view, ignoring the underlying structure of the scene. As illustrated in Fig. 2 (a) and (b), the Gaussian primitives fail to align with the surfaces. To overcome this limitation, we propose a seed-guided mechanism to control the distribution of 2D Gaussians.

(a) 3DGS (b) 2DGS (c) Ours

Fig. 2: Ground truth scene surface and Gaussian primitives distribution.

Seed-Guided Initialization. Starting from an SfM-derived point cloud $\mathbf{P} \in \mathbb{R}^{M\times 3}$, we apply voxelization to generate a set of seed points $\mathbf{V} \in \mathbb{R}^{N\times 3}$ by selecting the center points of each voxel grid to represent the seed points:

$$\mathbf{V} = \left\{ \left\lfloor \frac{\mathbf{P}}{\delta} \right\rfloor \cdot \delta \right\}, \tag{8}$$

where δ denotes the voxel size. Each seed point $v \in \mathbf{V}$ serves as the anchor for generating multiple 2D Gaussians whose positions are determined by learnable offsets. This initialization promotes a geometry-aware distribution of Gaussians and improves reconstruction robustness.

For each seed point $v \in \mathbf{V}$, we initialize a set of k Gaussians $\mathcal{G}_{i,j}$, where the position of each Gaussian is defined by:

$$\mathbf{p}_{i,j} = \mathbf{v}_i + \mathbf{O}_{i,j}, \tag{9}$$

with $\mathbf{p}_{i,j} \in \mathbb{R}^3$ denoting the global Gaussian position and $\mathbf{O}_{i,j} \in \mathbb{R}^3$ being a learnable offset optimized during training to align Gaussians with local geometry.

Each 2D Gaussian is further parameterized by scaling $\mathbf{s} \in \mathbb{R}^2$, rotation $\mathbf{t} \in \mathbb{R}^2$, color $\mathbf{c} \in \mathbb{R}^3$, and opacity $\alpha \in \mathbb{R}$. At initialization, scaling and rotation are estimated from local geometry derived from the point cloud, providing a spatially consistent starting point. All parameters are subsequently refined through optimization to achieve accurate alignment and representation.

Seed-Guided Optimization. To adaptively capture varying levels of detail in indoor scenes, we dynamically adjust the density of seed points via a combination of gradient-guided growth and contribution-based pruning.

We employ a gradient-guided growth strategy to densify seeds in structurally complex regions. For each voxel, we compute the average gradient ∇v of 2D Gaussians over N_g iterations. When ∇v exceeds a threshold θ_g, new seeds are introduced. This process operates within a multi-resolution voxel structure using adaptive thresholds, ensuring higher seed density in detail-rich areas.

Moreover, we implement contribution-based pruning to remove low-impact seeds. For each seed, We calculate the cumulative opacity α_v of connected 2D Gaussians over N_α iterations. If α_v falls below θ_α, the seed is pruned due to its

negligible contribution. This allocates Gaussians to structurally significant regions, optimizing both computational efficiency and reconstruction quality.

B. Monocular Cues Supervision

While the control of seed points enhances the structural consistency of the scene, it remains insufficient for achieving highly accurate geometry, particularly in detailed or textureless regions which are common in indoor environments. Therefore, we incorporate depth and surface normal priors, providing geometric constraints to further improve the scene reconstruction.

Monocular Depth Supervision. To enhance spatial alignment, we supervise the rendered depth $\hat{\mathcal{D}}$ using reference depths $\mathcal{D}$ predicted by a pre-trained model [30], via a scale-and-shift-invariant loss. Optimal scale s and shift t are computed to align $\hat{\mathcal{D}}$ with $\mathcal{D}$, producing $\hat{\mathcal{D}}_{\text{aligned}} = s \cdot \hat{\mathcal{D}} + t$. The depth loss $\mathcal{L}_d$ includes an MSE term between $\hat{\mathcal{D}}_{\text{aligned}}$ and $\mathcal{D}$, and a gradient regularization term $\mathcal{L}_{\text{grad}}$ for local smoothness:

$$\mathcal{L}_d = \frac{1}{|\mathcal{V}_d|} \sum \|\hat{\mathcal{D}}_{\text{aligned}} - \mathcal{D}\|^2 + \lambda_{grad} \cdot \mathcal{L}_{\text{grad}}, \tag{10}$$

where $|\mathcal{V}_d|$ denotes the number of valid pixels.

Monocular Normal Supervision. To improve reconstruction in textureless or geometrically regular regions (e.g., walls and floors), we impose normal supervision using predicted surface normals $\hat{\mathcal{N}}$ from a pre-trained model [33]. Let $\mathcal{N}$ denote the rendered normals. We formulate two objectives.

First, an $\mathcal{L}1$ loss minimizes the per-pixel difference between $\mathcal{N}$ and $\hat{\mathcal{N}}$:

$$\mathcal{L}_1 = \frac{1}{|\mathcal{V}_n|} \sum \left| \mathcal{N} - \hat{\mathcal{N}} \right|, \tag{11}$$

where $|\mathcal{V}_n|$ is the number of valid pixels.

Second, to enforce angular alignment, we adopt a cosine similarity loss that penalizes deviations in orientation:

$$\mathcal{L}_{\cos} = \frac{1}{|\mathcal{V}_n|} \sum \left(1 - \frac{\mathcal{N} \cdot \hat{\mathcal{N}}}{|\mathcal{N}| \cdot |\hat{\mathcal{N}}|} \right). \tag{12}$$

C. Multi-View Consistency Constraints

While the strategies above substantially improve reconstruction accuracy, small floaters may still persist due to lighting variations and subtle structures common in indoor scenes. To further refine the geometry, we introduce multi-view consistency constraints that reduce view-dependent inconsistencies. As illustrated in Fig. 1, for a reference view V_r, we select a neighboring view V_n and enforce both geometric and photometric consistency.

Geometric Consistency. We define a pixel-wise geometric loss that penalizes inconsistencies in forward and backward projections between views. Specifically, the homography H_{rn} that maps a homogeneous pixel coordinate $\tilde{\mathbf{p}}_r$ in the reference view V_r to the neighboring view V_n is defined as:

$$H_{rn} = K_n \left(R_{rn} - \frac{T_{rn} \mathbf{n}_r^\top}{d_r} \right) K_r^{-1}, \tag{13}$$

where K_r and K_n are the intrinsic matrices, R_{rn} and T_{rn} denote the relative pose from V_r to V_n, and $\mathbf{n}_r$, d_r are the surface normal and depth at $\mathbf{p}_r$.

The consistency is enforced by projecting $\tilde{\mathbf{p}}_r$ to V_n via H_{rn}, then back to V_r via H_{nr}. The geometric loss is:

$$\mathcal{L}_{geo} = \frac{1}{|\mathcal{V}_e|} \sum_{\mathbf{p}_r \in \mathcal{V}_e} \|\tilde{\mathbf{p}}_r - H_{nr} H_{rn} \tilde{\mathbf{p}}_r\|_2, \tag{14}$$

where $\mathcal{V}_e$ is the set of valid pixels with reliable projections.

Photometric Consistency Constraint. To account for local texture and illumination variations, we also enforce photometric consistency using normalized cross-correlation (NCC) [34], which measures similarity between image patches across views. Specifically, we convert color images to grayscale, and define the photometric loss as:

$$\mathcal{L}_{pho} = \frac{1}{|\mathcal{V}_e|} \sum_{\mathbf{p}_r \in \mathcal{V}_e} \left(1 - \text{NCC}(P_r(\tilde{\mathbf{p}}_r), P_n(H_{rn} \tilde{\mathbf{p}}_r))\right), \tag{15}$$

where P_r and P_n denote grayscale patches centered at $\tilde{\mathbf{p}}_r$ in the reference and neighboring views, respectively, and H_{rn} is the homography from V_r to V_n.

Finally, the multi-view consistency loss $\mathcal{L}_{mv}$ is given by:

$$\mathcal{L}_{mv} = \lambda_{geo} \mathcal{L}_{geo} + \lambda_{pho} \mathcal{L}_{pho}. \tag{16}$$

D. Optimization

In summary, with $\mathcal{L}_{rgb}$ representing the photometric supervision that minimizes the difference between rendered and input images proposed in the original 2DGS, our final training loss $\mathcal{L}$ is given by:

$$\mathcal{L} = \mathcal{L}_{rgb} + \lambda_d \cdot \mathcal{L}_d + \lambda_n \cdot \mathcal{L}_n + \mathcal{L}_{mv}. \tag{17}$$

TABLE I: Quantitative reconstruction comparison on ScanNet and ScanNet++ dataset.

Method	ScanNet [31]					ScanNet++ [32]				
	Acc. ↓	Comp. ↓	Prec. ↑	Recall ↑	F-score ↑	Acc. ↓	Comp. ↓	Prec. ↑	Recall ↑	F-score ↑
NeuS [18]	0.105	0.124	0.448	0.378	0.409	0.160	0.224	0.294	0.221	0.251
Neuralangelo [19]	0.185	0.223	0.252	0.260	0.255	0.363	0.264	0.172	0.120	0.141
3DGS [9]	0.338	0.406	0.129	0.067	0.085	0.144	0.990	0.322	0.066	0.104
SuGaR [26]	0.167	0.148	0.361	0.373	0.366	0.158	0.178	0.383	0.349	0.361
2DGS [10]	0.157	0.151	0.336	0.347	0.341	0.359	0.228	0.230	0.160	0.183
PGSR [27]	0.125	0.117	0.420	0.433	0.426	0.204	0.202	0.353	0.217	0.249
RaDe-GS [28]	0.167	0.205	0.309	0.307	0.306	0.284	0.252	0.171	0.179	0.166
2DGS-Room (Ours)	0.055	0.092	0.648	0.518	0.575	0.262	0.112	0.450	0.498	0.464

V. Experiments

A. Experimental Setup

Dataset. We evaluate the performance of our approach on reconstruction quality across 12 real-world indoor scenes randomly selected from publicly available datasets: 8 scenes from ScanNet(V2) [31] and 4 scenes from ScanNet++ [32].

Implementation Details. Our training strategy and hyperparameters are consistent with the baseline 2DGS method. We set $\lambda_{geo} = 0.05$, $\lambda_{pho} = 0.2$, $\lambda_d = \lambda_n = 1.0$. We render depth maps for all training views and then adopt TSDF fusion [35] for mesh extraction. We train all models for 30k iterations. All experiments are conducted on an NVIDIA RTX 4090 GPU.

Metrics. Consistent with existing methods [5], [6], five metrics

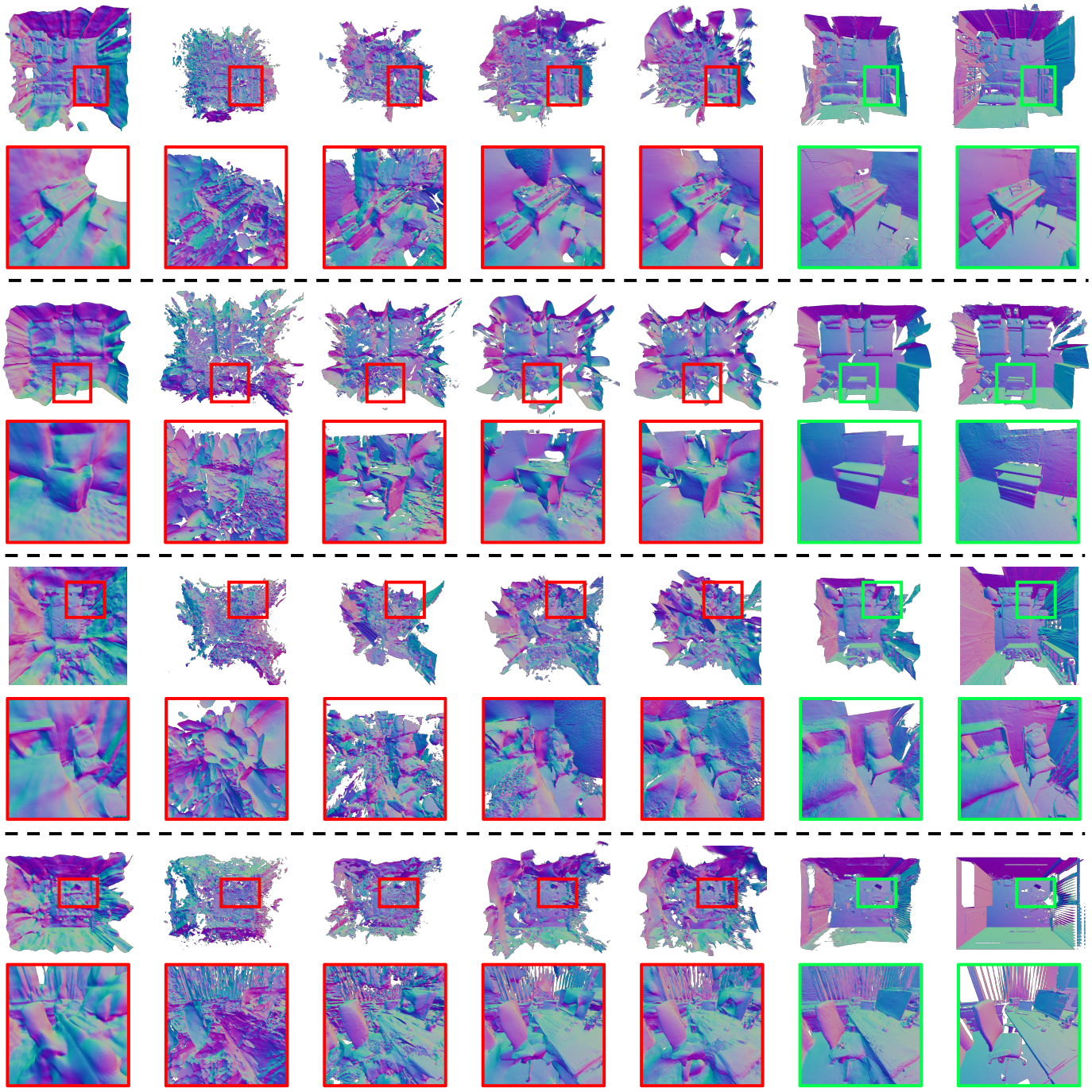

Fig. 3: Qualitative reconstruction comparisons. For each scene, the first row is the top view of the whole room, and the second row is the details of the masked region.

are employed to evaluate the quality of reconstructed meshes: Accuracy, Completion, Precision, Recall, and F-score.

Baselines. We compare our approach with several state-of-the-art methods, covering both neural volume rendering and Gaussian splatting techniques. The baselines include (1) Neural volume rendering methods: NeuS [18] and NeuralAngelo [19]; (2) Gaussian splatting methods: 3DGS [9], SuGaR [26], RaDe-GS [28], PGSR [27], and 2DGS [10].

B. Results Analysis

Qualitative Results. To show the visualized reconstruction results of our method, we compare our 2DGS-Room with different reconstruction methods and the ground truth. As shown in Figure 3, the overall quality of our reconstructions is noticeably higher. In comparison with Gaussian-based methods, our method obtains a more visually coherent and accurate representation of the indoor scenes, with well-defined surfaces and consistent details across different views.

Quantitative Results. Quantitative results are presented in Table I, showing a comprehensive comparison in geometry metrics on indoor scene datasets. On the ScanNet dataset, our method achieves the best results in all metrics. Compared to NeRF-based methods [18], [19] which typically require over 20 hours to train a scene, our method significantly reduces training time, being approximately 30 times faster.

C. Ablation Studies

To assess the individual contributions of each component in our model, we perform ablation studies on the ScanNet dataset. The qualitative results are shown in Figure 4 and Table II reports the quantitative results.

Seed Points Guidance. Without guidance, scenes lack structure and appear inflated. Adding seed points enables our method to better capture the underlying geometric framework, boosting the F-score by 87.3%.

Monocular Depth. Removing depth supervision causes spatial misalignment. Incorporating it significantly enhances geometric accuracy, increasing the F-score by 31.3%.

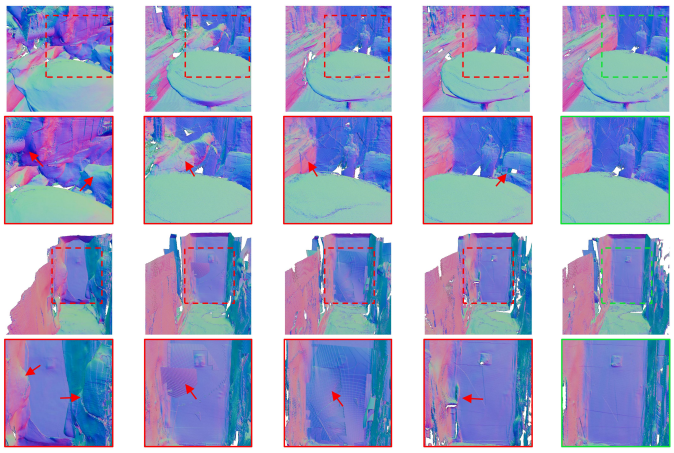

Fig. 4: Qualitative results of ablation study.

TABLE II: Results of the ablation study on ScanNet dataset.

Method	Acc.↓	Comp.↓	Prec.↑	Recall↑	F-score↑
w/o Seed	0.128	0.152	0.336	0.284	0.307
w/o Depth	0.084	0.139	0.510	0.386	0.438
w/o Normal	0.066	0.102	0.596	0.463	0.520
w/o MV	0.055	0.092	0.644	0.508	0.566
Full model	**0.055**	**0.092**	**0.648**	**0.518**	**0.575**

Monocular Normal. Lack of normal priors leads to inconsistent planar surfaces (e.g., walls). Adding supervision improves surface alignment, increasing the F-score by 10.6%.

Multi-View Consistency. Without this constraint, Gaussians fail to align spatially. This module mitigates view-dependent inconsistencies and further refines reconstruction quality.

VI. Conclusion

We propose 2DGS-Room, a novel framework for indoor scene reconstruction based on 2D Gaussian Splatting. Our method introduces a seed-guided optimization strategy that leverages structural information to direct local Gaussian distributions. By integrating geometric priors and multi-view consistency, we significantly enhance reconstruction quality in textureless regions and complex details while mitigating view-dependent artifacts. Extensive experiments demonstrate that our method outperforms existing techniques across multiple metrics and diverse indoor scenes.

VII. Acknowledgment

This paper was supported by the Shenzhen Science and Technology Major Special Project (KJZD20230923115503007).

References

[1] Y.-F. Tang, C. Tai, F.-X. Chen, W.-T. Zhang, T. Zhang, X.-P. Liu, Y.-J. Liu, and L. Zeng, "Mobile robot oriented large-scale indoor dataset for dynamic scene understanding," in *ICRA*. IEEE, 2024, pp. 613–620.

[2] X. H. Jing, X. Xiong, F. H. Li, T. Zhang, and L. Zeng, "A two-stage reinforcement learning approach for robot navigation in long-range indoor dense crowd environments," in *IROS*. IEEE, 2024, pp. 5489–5496.

[3] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *ECCV*. Springer, 2016, pp. 501–518.

[4] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *ECCV*, 2018, pp. 767–783.

[5] J. Wang, P. Wang, X. Long, C. Theobalt, T. Komura, L. Liu, and W. Wang, "Neuris: Neural reconstruction of indoor scenes using normal priors," in *ECCV*. Springer, 2022, pp. 139–155.

[6] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger, "Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction," *Advances in neural information processing systems*, vol. 35, pp. 25 018–25 032, 2022.

[7] X. Li, Y. Ding, J. Guo, X. Lai, S. Ren, W. Feng, and L. Zeng, "Edge-aware neural implicit surface reconstruction," in *ICME*. IEEE, 2023, pp. 1643–1648.

[8] X. Li, Y. Ji, X. Lai, W. Zhang, and L. Zeng, "Fine-detailed neural indoor scene reconstruction using multi-level importance sampling and multi-view consistency," in *ICIP*. IEEE, 2024, pp. 3477–3483.

[9] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.

[10] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao, "2d gaussian splatting for geometrically accurate radiance fields," in *ACM SIGGRAPH 2024 conference papers*, 2024, pp. 1–11.

[11] A. Broadhurst, T. W. Drummond, and R. Cipolla, "A probabilistic framework for space carving," in *Proceedings eighth IEEE international conference on computer vision*, vol. 1. IEEE, 2001, pp. 388–393.

[12] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[13] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," in *ICCV*, 2021, pp. 5855–5864.

[14] K. Deng, A. Liu, J.-Y. Zhu, and D. Ramanan, "Depth-supervised nerf: Fewer views and faster training for free," in *CVPR*, 2022, pp. 12 882–12 891.

[15] M. Niemeyer, J. T. Barron, B. Mildenhall, M. S. Sajjadi, A. Geiger, and N. Radwan, "Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs," in *CVPR*, 2022, pp. 5480–5490.

[16] G. Wang, Z. Chen, C. C. Loy, and Z. Liu, "Sparsenerf: Distilling depth ranking for few-shot novel view synthesis," in *ICCV*, 2023, pp. 9065–9076.

[17] M. Oechsle, S. Peng, and A. Geiger, "Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction," in *ICCV*, 2021, pp. 5589–5599.

[18] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *NeurIPS*, 2021.

[19] Z. Li, T. Müller, A. Evans, R. H. Taylor, M. Unberath, M.-Y. Liu, and C.-H. Lin, "Neuralangelo: High-fidelity neural surface reconstruction," in *CVPR*, 2023, pp. 8456–8465.

[20] Q. Fu, Q. Xu, Y. S. Ong, and W. Tao, "Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3403–3416, 2022.

[21] H. Guo, S. Peng, H. Lin, Q. Wang, G. Zhang, H. Bao, and X. Zhou, "Neural 3d scene reconstruction with the manhattan-world assumption," in *CVPR*, 2022, pp. 5511–5520.

[22] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *CVPR*, 2016, pp. 4104–4113.

[23] M. Turkulainen, X. Ren, I. Melekhov, O. Seiskari, E. Rahtu, and J. Kannala, "Dn-splatter: Depth and normal priors for gaussian splatting and meshing," *WACV*, 2025.

[24] H. Xiang, X. Li, X. Lai, W. Zhang, Z. Liao, K. Cheng, and X. Liu, "Gaussianroom: Improving 3d gaussian splatting with sdf guidance and monocular cues for indoor scene reconstruction," in *ICRA*. IEEE, 2025.

[25] M. Yu, T. Lu, L. Xu, L. Jiang, Y. Xiangli, and B. Dai, "Gsdf: 3dgs meets sdf for improved rendering and reconstruction," *arXiv preprint arXiv:2403.16964*, 2024.

[26] A. Guédon and V. Lepetit, "Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering," *CVPR*, 2024.

[27] D. Chen, H. Li, W. Ye, Y. Wang, W. Xie, S. Zhai, N. Wang, H. Liu, H. Bao, and G. Zhang, "Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction," *arXiv preprint arXiv:2406.06521*, 2024.

[28] B. Zhang, C. Fang, R. Shrestha, Y. Liang, X. Long, and P. Tan, "Rade-gs: Rasterizing depth in gaussian splatting," *arXiv preprint arXiv:2406.01467*, 2024.

[29] T. Lu, M. Yu, L. Xu, Y. Xiangli, L. Wang, D. Lin, and B. Dai, "Scaffold-gs: Structured 3d gaussians for view-adaptive rendering," in *CVPR*, 2024, pp. 20 654–20 664.

[30] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun, "Depth pro: Sharp monocular metric depth in less than a second," *arXiv preprint arXiv:2410.02073*, 2024.

[31] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *CVPR*, 2017, pp. 5828–5839.

[32] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai, "Scannet++: A high-fidelity dataset of 3d indoor scenes," in *ICCV*, 2023, pp. 12–22.

[33] G. Bae, I. Budvytis, and R. Cipolla, "Estimating and exploiting the aleatoric uncertainty in surface normal estimation," in *ICCV*, 2021, pp. 13 137–13 146.

[34] J.-C. Yoo and T. H. Han, "Fast normalized cross-correlation," *Circuits, systems and signal processing*, vol. 28, pp. 819–843, 2009.

[35] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *ACM SIGGRAPH*, 1996, pp. 303–312.