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

Wanting Zhang   Haodong Xiang   Zhichao Liao   Xiansong Lai   Xinghui Li[†]   Long Zeng[†]
Tsinghua University
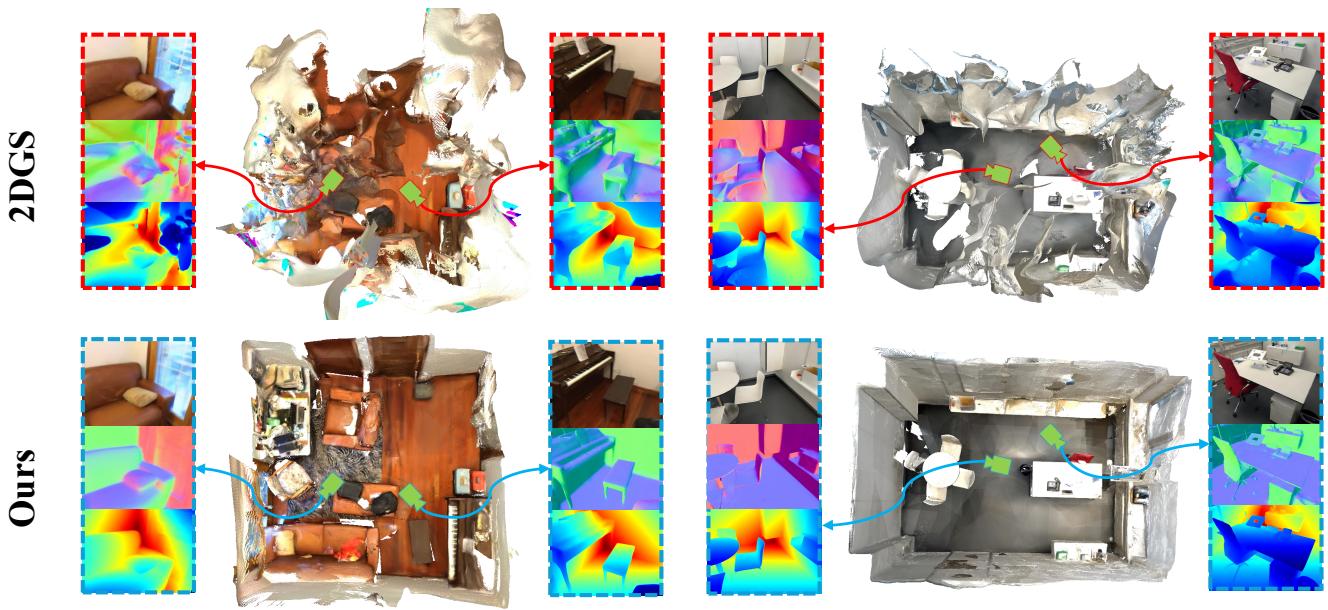https://valentina-zhang.github.io/2DGS-Room/

Figure 1. **2DGS-Room achieves high-fidelity geometric reconstructions for indoor scenes.** We introduce seed points to guide the distribution of 2D Gaussians coupled with geometric constraints, leading to clearer structures and more accurate geometry.

Abstract

*The reconstruction of indoor scenes remains challenging due to the inherent complexity of spatial structures and the prevalence of textureless regions. Recent advancements in 3D Gaussian Splatting have improved novel view synthesis with accelerated processing but have yet to deliver comparable performance in surface reconstruction. In this paper, we introduce **2DGS-Room**, a novel method leveraging 2D Gaussian Splatting for high-fidelity indoor scene reconstruction. Specifically, we employ a seed-guided mechanism to control the distribution of 2D Gaussians, with the density of seed points dynamically optimized through adaptive growth and pruning mechanisms. To further improve geometric accuracy, we incorporate monocular depth and normal priors to provide constraints for details and textureless regions respectively. Additionally, multi-view consistency constraints are employed to mitigate artifacts and further enhance reconstruction quality. Extensive experiments on ScanNet and ScanNet++ datasets demonstrate that our method achieves state-of-the-art performance in indoor scene reconstruction.*

1. Introduction

3D reconstruction from multi-view RGB images is a fundamental task in the fields of computer vision and computer graphics. The reconstructed models can be utilized in a wide range of applications, including virtual reality, video games, autonomous driving, and robotics. Reconstructing indoor scenes is a challenging task in the field of 3D reconstruction, as indoor environments often contain large textureless regions. MVS-based methods [1–3] often yield incomplete or geometrically flawed reconstructions, primarily due to the geometric ambiguities arising from the presence of textureless regions.

Recent advancements in neural-radiance-field-based methods [4–8] that utilize signed distance fields (SDF) for scene modeling have enabled accurate and complete mesh reconstruction in indoor environments. This progress is attributed to the continuity of neural SDFs and the integration of monocular geometric priors [6]. Although neural-radiance-field-based methods achieve high-quality reconstruction, they are computationally expensive due to the need for dense ray sampling, resulting in long optimization times. Fortunately, 3D Gaussian Splatting (3DGS) [9] enhances the optimization and rendering efficiency of neural rendering through its differentiable rasterization technique, offering new possibilities for 3D scene reconstruction. 2DGS [10] build upon 3DGS by using 2D-oriented planar Gaussians as primitives, significantly improving surface reconstruction quality. Despite these advances, Gaussian splatting-based methods still often produce floating artifacts and incomplete reconstructions in indoor scenes, due to the lack of structured geometric constraints.

In this work, we present a novel approach named **2DGS-Room**, aiming to achieve high-fidelity geometric reconstruction for indoor scenes based on 2D Gaussian Splatting. Considering the scene's underlying structure, we propose a seed-guided mechanism to control the distribution and density of 2D Gaussians. Specifically, we introduce a seed-guided initialization to generate 2D Gaussians, ensuring their alignment with scene surfaces to improve geometric accuracy. To further refine the reconstruction, we propose a seed-guided optimization strategy that dynamically adjusts seed point density through gradient-guided growth and contribution-based pruning, enabling efficient representation of fine details. Additionally, we incorporate monocular depth and normal priors to provide crucial geometric constraints. The depth prior addresses distortions in detailed areas, while the normal prior ensures accurate surface estimation in textureless regions. Furthermore, we introduce multi-view consistency constraints to address residual artifacts, which enforces both geometric and photometric consistency across multiple views.

Extensive qualitative and quantitative experiments show that compared with Gaussian-based methods, 2DGS-Room achieves start-of-the-art performance in indoor scenarios. 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.

2. Related work

2.1. Multi-View Stereo

Multi-view stereo (MVS) methods [1, 11–13] estimate the 3D coordinates of pixels and explicitly reconstruct objects and scenes by matching features across a collection of posed images. The surface is then obtained through the application of Poisson surface reconstruction [14]. In indoor scenes, particularly in large texture-less regions, these methods frequently encounter difficulties due to the scarcity of features. Voxel-based approaches [15–18] optimize spatial occupancy and color within a voxel grid, thus avoiding the challenges of feature matching. However, high-resolution memory constraints degrade reconstruction quality. Learning-based multi-view stereo methods [2, 3, 19–25] implicitly match corresponding multi-view features through neural networks, enabling end-to-end 3D reconstruction. Nonetheless, even with extensive training data, errors may still occur in the results when handling occlusions, complex lighting, or regions with subtle textures.

2.2. Neural Radiance Field

Neural Radiance Fields (NeRF) [26] 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 [27–29] enhance rendering quality by improving the ray sampling strategy. Other works [30–34] accelerate training and rendering through techniques such as multi-resolution hash encoding or resizing MLPs. Some studies aim to enhance rendering quality by incorporating regularization terms. For example, depth regularization [35, 36] explicitly supervises ray termination to minimize unnecessary sampling time. Other approaches focus on enforcing smoothness constraints on rendered depth maps [37] or utilizing multi-view consistency regularization in sparse-view scenarios [38, 39]. Some research explores the use of alternative implicit functions to enhance the geometric reconstruction capabilities of NeRF, such as occupancy grids [40, 41] and signed distance functions (SDFs) [4, 5, 34, 42, 43], replacing NeRF's volumetric density field. To further enhance reconstruction quality, [44, 45] suggest regularizing optimization with SfM points, while [6, 46] incorporate priors like the Manhattan world assumption and pseudo depth supervision. However, these approaches often lead to incomplete reconstructions and require extensive optimization time.
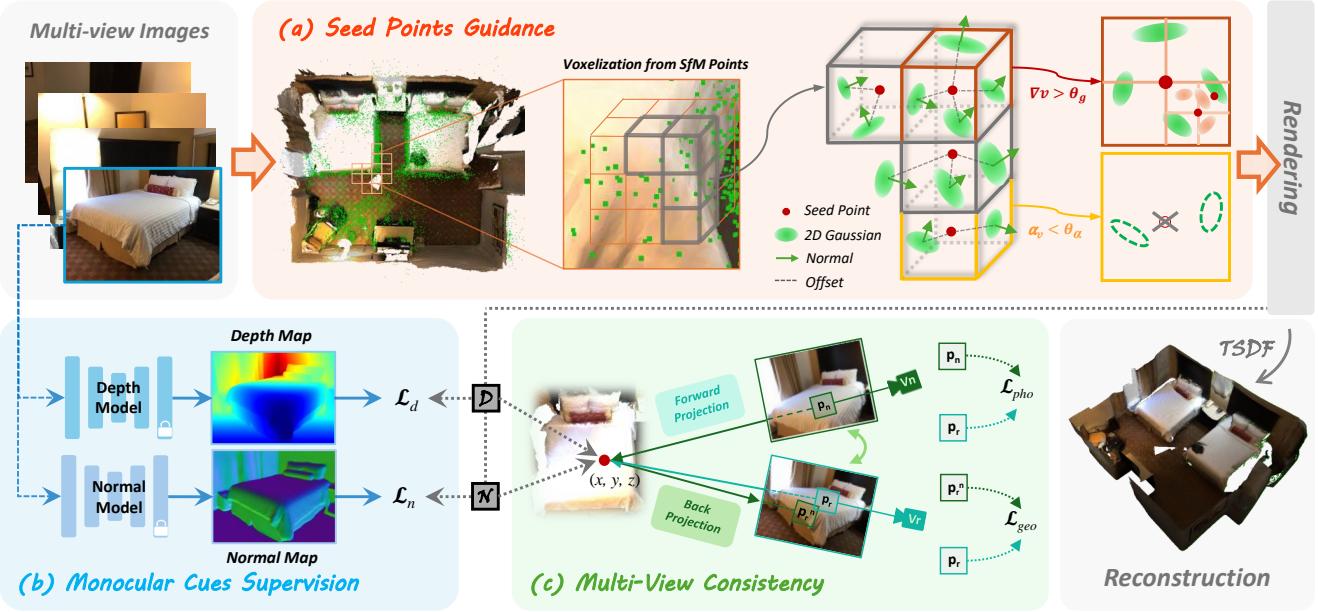
Figure 2. **Overview of 2DGS-Room.** Given multi-view posed images, we improve 2DGS to achieve high-fidelity geometric reconstruction for indoor scenes. (a) Starting from an SfM-derived point cloud, we generate a set of seed points through voxelization, establishing a stable foundation for guiding the distribution and density of 2D Gaussians. We further introduce an adaptive growth and pruning strategy to optimize seed points. (b) We incorporate depth and normal priors, addressing the challenges of detailed areas and textureless regions. (c) We introduce multi-view consistency constraints to further enhance the quality of the indoor scene reconstruction.

2.3. Gaussian Splatting

3D Gaussian Splatting [9] explicitly represents 3D scenes using learnable Gaussian primitives, enabling high-quality novel view synthesis with short training times and high rendering frame rates. The 3DGS method is solely responsible for the image loss, and after initializing with sparse point clouds generated by SfM [47], no further constraints are applied to the Gaussian primitives. This leads to a disorganized distribution of the optimized Gaussian primitives, resulting in poor geometric properties. Works such as DN-Splatter [48], GaussianRoom [49] and GSDF [50] introduce geometric priors or leverage the accurate geometric information from SDFs to supervise the optimization of Gaussians. SuGaR [51], PGSR [52] and RaDe-GS [53] use Flatten Gaussians to represent scenes, enhancing surface reconstruction capabilities. In contrast, 2DGS [10] directly applies 2D oriented planar Gaussians instead of 3D Gaussian primitives to represent 3D scenes, achieving better surface reconstruction results. However, it still encounters poor reconstruction in indoor scenes due to Gaussian primitives lacking geometric constraints.

3. Preliminary

The key innovation of 2DGS [10] lies in its transformation of 3D volumetric Gaussians into flat 2D Gaussians, or surfels, for scene representation. It directly models scenes with 2D elliptical disks, simplifying the representation process and yielding more accurate geometry without extra mesh refinement.

Each 2D Gaussian disk, defined in a local tangent plane, is parameterized by a central point $\mathbf{p}_k$, two orthogonal tangential vectors $\mathbf{t}_u$ and $\mathbf{t}_v$, and a scaling vector (s_u, s_v) that controls the variances along each direction. The normal $\mathbf{t}_w$ of each Gaussian disk is computed as $\mathbf{t}_w = \mathbf{t}_u \times \mathbf{t}_v$ and this orientation can be arranged into a rotation matrix $\mathbf{R} = [\mathbf{t}_u, \mathbf{t}_v, \mathbf{t}_w]$. The scaling factors can be arranged into a 3×3 diagonal matrix $\mathbf{S} = [s_u, s_v, 0]$. Then a 2D Gaussian can be parameterized:

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

where $\mathbf{H} \in 4 \times 4$ is a homogeneous transformation matrix representing the geometry of the 2D Gaussian:

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

In the Gaussian's tangent frame (u, v), the 2D Gaussian value $\mathcal{G}(\mathbf{u})$ at point $\mathbf{u} = (u, v)$ is evaluated as:

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

For efficient rendering, each 2D Gaussian is projected onto the image plane by a general 2D-to-2D mapping in ho-

mogeneous coordinates. Given a world-to-screen transformation matrix $\mathbf{W}$, the screen space points can be derived from:

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

where $\mathbf{x}$ represents a homogeneous ray emitted from the camera and passing through pixel (x, y) and intersecting the splat at depth z.

To avoid numerical instability, a ray-splat intersection is calculated explicitly by finding the intersection of three non-parallel planes in the 3D scene. Given an image coordinate $\mathbf{x} = (x, y)$, the ray of a pixel can be defined by the intersection of two homogeneous planes: the x-plane $\mathbf{h}_x = (-1, 0, 0, x)$ and the y-plane $\mathbf{h}_y = (0, -1, 0, y)$. To compute the intersection with the Gaussian splat, both planes are transformed to uv-space:

$$\mathbf{h}_u = (\mathrm{WH})^\top \mathbf{h}_x, \quad \mathbf{h}_v = (\mathrm{WH})^\top \mathbf{h}_y. \qquad (5)$$

By homography, the two planes are used to find the intersection point $(u(x), v(x))$ with the 2D Gaussian splats, given by:

$$u(\mathbf{x}) = \frac{\mathbf{h}_u^2 \mathbf{h}_v^4 - \mathbf{h}_u^4 \mathbf{h}_v^2}{\mathbf{h}_u^1 \mathbf{h}_v^2 - \mathbf{h}_u^2 \mathbf{h}_v^1}, \quad v(\mathbf{x}) = \frac{\mathbf{h}_u^4 \mathbf{h}_v^1 - \mathbf{h}_u^1 \mathbf{h}_v^4}{\mathbf{h}_u^1 \mathbf{h}_v^2 - \mathbf{h}_u^2 \mathbf{h}_v^1}, \qquad (6)$$

where $\mathbf{h}_u^i$ and $\mathbf{h}_v^i$ are components of the transformed planes in the Gaussian's tangent frame.

4. Methods

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, which leverages seed points to control the distribution and density of 2D Gaussians, thereby improving the accuracy and efficiency of scene representation in indoor scenes (Sec. 4.1). To further improve geometric accuracy, we incorporate depth and normal priors, which enhance the representation of detailed areas and textureless regions, respectively (Sec. 4.2). Finally, to mitigate floating artifacts caused by lighting variations in indoor scenes, we introduce multi-view consistency constraints, further enhancing the quality of the indoor scene reconstruction (Sec. 4.3). An overview of our framework is provided in Fig. 2.

4.1. 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. 3 (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. Specifically, we utilize a set of seed points to provide a stable foundation for generating 2D Gaussians, ensuring that the reconstruction reflects

the underlying scene structure more accurately. Additionally, we introduce an adaptive growth and pruning strategy to dynamically adjust the density of seed points.

Seed-Guided Initialization. Starting from an SfM-derived point cloud $\mathbf{P} \in \mathbb{R}^{M \times 3}$, we first filter some unreliable outliers. We define a confidence measure $O_{\mathbf{p}_i}$ for each individual point $\mathbf{p}_i$ in the point cloud. This measure is expressed as follows:

$$O_{\mathbf{p}_i} = \begin{cases} 1 & \text{if } m \geq \epsilon \\ 0 & \text{if } m < \epsilon \end{cases}, \qquad (7)$$

where m represents the number of image feature matches associated with $\mathbf{p}_i$, and ϵ is a predefined threshold. Points with a number of matched features below ϵ are deemed unreliable and removed from the point cloud to ensure a more accurate reconstruction.

Following the filtering process, 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:

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

where δ denotes the voxel grid size. Each seed point $v \in \mathbf{V}$ serves as the basis for deriving several 2D Gaussians, which are positioned based on learnable offsets from the seed point. This initialization ensures that the distribution of Gaussians is closely aligned with the underlying geometry of the scene, thereby improving the overall robustness of the reconstruction quality.

For each seed point $v \in \mathbf{V}$, we initialize a set of k 2D Gaussians $\{\mathcal{G}_{i,j}\}$, where $\mathcal{G}_{i,j}$ denotes the j-th Gaussian associated with the i-th seed. The position of each Gaussian is determined by a learnable offset $\mathbf{O}_{i,j}$ from the seed point location:

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

where $\mathbf{p}_{i,j} \in \mathbb{R}^3$ represents the global position of the Gaussian, and $\mathbf{O}_{i,j} \in \mathbb{R}^3$ is a learnable offset which is optimized during training to adjust each Gaussian's local position for better alignment with the scene.

Expect for the center position, each 2D Gaussian is parameterized by the scaling $\mathbf{s} \in \mathbb{R}^2$, rotation $\mathbf{t} \in \mathbb{R}^2$, appearance $\mathbf{c} \in \mathbb{R}^3$ and opacity $\alpha \in \mathbb{R}$. At initialization, the scaling and rotation are aligned with the local geometry derived from the point cloud, which provides a starting approximation that reflects the scene's spatial distribution. During training, these parameters are iteratively optimized to refine the representation.

Seed-Guided Optimization. In order to capture different levels of detail in complex indoor scenes, we develop an adaptive approach to dynamically adjust seed point density by combining gradient-guided growth and contribution-based pruning.

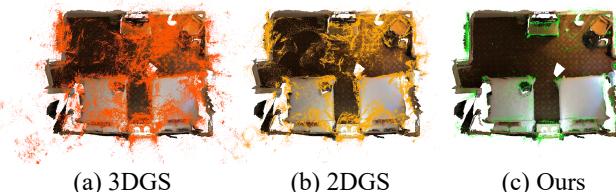|		
(a) 3DGS	(b) 2DGS	(c) Ours

Figure 3. **Ground truth scene surface and Gaussian primitives distribution.** Compared with 3DGS and 2DGS, our method significantly reduces scattered floaters in the non-surface areas, benefitting from our designed structured geometric constraints.

We utilize a gradient-guided growth strategy to increase seed point density adaptively, especially in areas with high structural complexity or fine details. For each voxel, we compute the average gradient ∇v of the included 2D Gaussians across N_g training iterations, using it as an indicator of structural complexity. When ∇v exceeds a threshold θ_g, additional seed points are introduced to enhance representation. This growth occurs within a multi-resolution voxel structure, with thresholds that adapt according to the resolution level, ensuring a higher seed density in regions requiring more detail.

Moreover, we implement a contribution-based pruning strategy that selectively removes low-impact seed points. For each seed, we calculate the cumulative opacity α_v of the connected 2D Gaussians over N_α iterations. If α_v is below a predefined threshold θ_α, the seed point is pruned, as its minimal contribution to scene opacity suggests the limited impact on the overall representation. This strategy allows us to allocate Gaussians to regions of higher structural significance, enhancing both computational efficiency and reconstruction quality.

4.2. 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. The depth prior is leveraged to mainly refine the spatial alignment of objects in the scene by aligning the rendered depths with reference depths predicted from a pre-trained model [54]. We incorporate depth supervision by aligning the rendered depths with reference depths through a scale-and-shift-invariant loss [55], compensating for relative scaling discrepancies that may arise in the representation of complex indoor geometries.

Given the rendered depths $\hat{\mathcal{D}}$, we first compute optimal scale s and shift t values to minimize discrepancies in scale and translation between our rendered depths and the

reference depths to address potential inconsistencies that may arise due to relative scaling differences in complex scenes. Then we adjust the predicted depth map to obtain the aligned prediction: $\hat{\mathcal{D}}_{\text{aligned}} = s \cdot \hat{\mathcal{D}} + t$.

The depth loss $\mathcal{L}_d$ consists of two terms: a data term that minimizes the mean squared error (MSE) between the aligned rendered depths $\hat{\mathcal{D}}_{\text{aligned}}$ and the reference depths $\mathcal{D}$, and a regularization term for gradient consistency that encourages local smoothness in the depth rendering. Formally, the depth loss is defined as:

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

where $|\mathcal{V}_d|$ represents the number of pixels with valid depths, and $\mathcal{L}_{\text{grad}}$ is a spatial regularization term that penalizes abrupt depth variations across neighboring pixels.

Monocular Normal Supervision. Additionally, the normal prior plays a crucial role in addressing the reconstruction challenges of textureless or planar regions like walls and floors. So we also incorporate normal supervision to enforce a smooth and realistic surface orientation throughout the scene.

Let $\hat{\mathcal{N}}$ denote the reference normals derived from a pretrained model [56], and $\mathcal{N}$ represents the rendered normals. We first use the $\mathcal{L}_1$ norm loss to quantify the absolute difference in magnitude between the rendered and reference normals, promoting consistency in the length of the vectors:

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

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

To further encourage the alignment of with $\hat{\mathcal{N}}$, we use a cosine similarity loss that penalizes angular differences between the two normal vectors:

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

The final normal supervision loss $\mathcal{L}_n$ is defined as:

$$\mathcal{L}_n = \lambda_1 \cdot \mathcal{L}_1 + \lambda_{\cos} \cdot \mathcal{L}_{\cos}. \quad (13)$$

4.3. Multi-View Consistency Constraints

The strategies outlined above significantly improve the accuracy of indoor scene reconstruction, but we observe that some small floaters may still persist in certain scenarios. These cases are likely caused by the complex lighting variations and subtle spatial structures typical in indoor environments. Therefore, we introduce multi-view consistency constraints to further refine the reconstruction by reducing the inconsistencies that occasionally manifest across different views. Specifically, as shown in Figure 2, given a

reference view V_r, we select a neighboring view V_n and enforce geometric consistency and photometric consistency between the two views.

Geometric Consistency Constraint. To ensure consistent geometry across views, we define a pixel-wise geometric consistency loss that penalizes discrepancies in the forward and backward projections for each individual pixel.

We compute a transformation H_{rn} to represent the homography matrix mapping a pixel $\mathbf{p}_r$ from V_r to the corresponding pixel $\mathbf{p}_n$ in V_n:

$$H_{rn} = K_n \left(R_{rn} - \frac{T_{rn}\mathcal{N}_r^\top}{\mathcal{D}_r} \right) K_r^{-1}, \qquad (14)$$

where K denotes the camera's intrinsic matrix. R_{rn} and T_{rn} are the relative rotation and translation from the reference frame to the neighboring frame.

For each pixel $\mathbf{p}_r$, we project it forward from V_r to V_n using H_{rn}, and then back-project from V_n to V_r using H_{nr}. The resulting multi-view geometric consistency loss $\mathcal{L}_{\text{geo}}$ is formulated as:

$$\mathcal{L}_{geo} = \frac{1}{|\mathcal{V}_e|} \sum_{\mathbf{p}_r \in \mathcal{V}_e} \|\mathbf{p}_r - H_{nr}H_{rn}\mathbf{p}_r\|, \qquad (15)$$

where $\mathcal{V}_e$ is a set of valid pixels excluding those with high forward and backward projection errors.

Photometric Consistency Constraint. To account for local variations in texture and illumination, we also enforce photometric consistency which is measured using the normalized cross-correlation (NCC) [59], penalizing differences in pixel intensity distributions between the views.

Focusing on geometric details, we convert color images into grayscale and the photometric consistency loss $\mathcal{L}_{pho}$ is defined as:

$$\mathcal{L}_{pho} = \frac{1}{|\mathcal{V}_e|} \sum_{\mathbf{p}_r \in \mathcal{V}_e} \left(1 - \text{NCC}(G_r(\mathbf{p}_r), G_n(H_{rn}\mathbf{p}_r))\right), \qquad (16)$$

where G_r and G_n denote the grayscale intensities of the patches in V_r and V_n, respectively.

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

$$\mathcal{L}_{mv} = \lambda_{geo}\mathcal{L}_{geo} + \lambda_{pho}\mathcal{L}_{pho}. \qquad (17)$$

4.4. 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:

$$\mathcal{L} = \mathcal{L}_{rgb} + \lambda_d \cdot \mathcal{L}_d + \lambda_n \cdot \mathcal{L}_n + \mathcal{L}_{mv}, \qquad (18)$$

where λ_d and λ_n control the relative contributions of depth and normal supervision, respectively.

5. Experiments

5.1. Experimental Setup

Dataset. We evaluate the performance of our approach on reconstruction quality across 12 real-world indoor scenes from publicly available datasets: 8 scenes from Scan-Net(V2) [57] and 4 scenes from ScanNet++ [58].

Implementation Details. Our training strategy and hyperparameters are consistent with the baseline 2DGS method to ensure comparability. We set $k = 10$, $\lambda_1 = 0.01$, $\lambda_{\cos} = 0.01$, $\lambda_{grad} = 0.5$, $\lambda_{geo} = 0.05$, $\lambda_{pho} = 0.2$, $\lambda_d = 1.0$, $\lambda_n = 1.0$, in all our experiments. We render depth maps for all training views and then adopt TSDF fusion [60] for mesh extraction. We train all models for 30k iterations. All experiments are conducted on an NVIDIA RTX 4090 GPU to ensure consistent processing.

Metrics. Consistent with existing methods [5, 6], five standard metrics are employed to evaluate the quality of reconstructed meshes: Accuracy, Completion, Precision, Recall, and F-score.

Method	ScanNet [57]					ScanNet++ [58]				
	Acc. ↓	Comp. ↓	Prec. ↑	Recall ↑	F-score ↑	Acc. ↓	Comp. ↓	Prec. ↑	Recall ↑	F-score ↑
NeuS [4]	0.105	0.124	0.448	0.378	0.409	0.160	0.224	0.294	0.221	0.251
Neuralangelo [34]	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 [51]	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 [52]	0.125	0.117	0.420	0.433	0.426	0.204	0.202	0.353	0.217	0.249
RaDe-GS [53]	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

Table 1. **Quantitative reconstruction comparison on ScanNet and ScanNet++ dataset.** Averaged results are reported over 8 scenes and 4 scenes, respectively. 2DGS-Room achieves the best F-score.
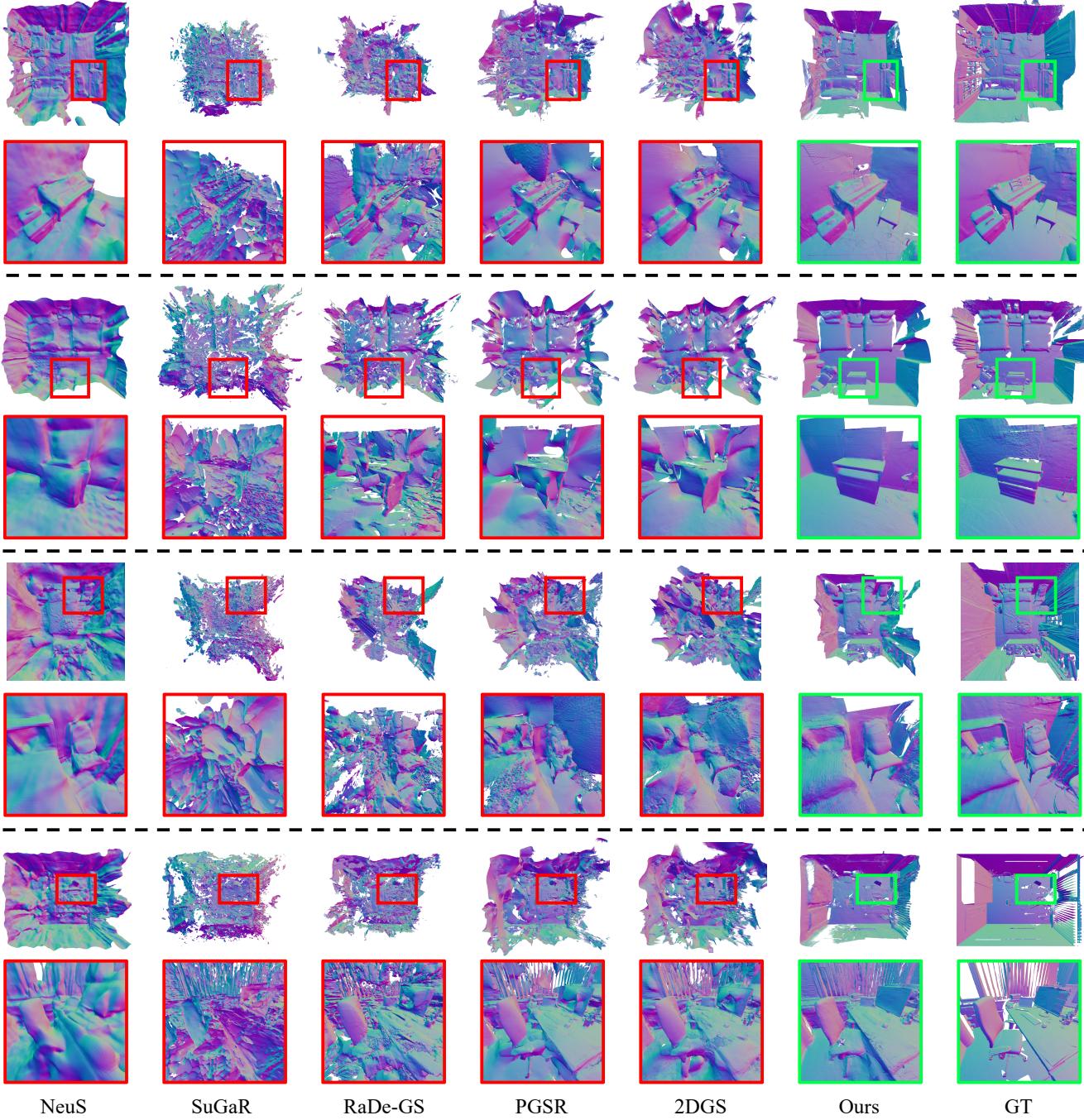
Figure 4. **Qualitative reconstruction comparisons.** For each indoor scene, the first row is the top view of the whole room, and the second row is the details of the masked region.

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 [4] and Neu-ralAngelo [34]; (2) Gaussian splatting methods: 3DGS [9], SuGaR [51], RaDe-GS [53], PGSR [52], and 2DGS [10].

5.2. Results Analysis

Qualitative Results. To show the visualized reconstruction results of our method, we compare our 2DGS-Room with different reconstruction methods, including NeuS [4], SuGaR [51], RaDe-GS [53], PGSR [52], 2DGS [10], and the ground truth. As illustrated in Figure 4, our method ex-

7

hibits significantly clearer scene structures, which is largely attributed to the seed-guided strategy. Additionally, thanks to the incorporation of depth and normal priors, 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 1, 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 [4, 34] which typically require over 20 hours to train a scene, our method significantly reduces training time, being approximately 30 times faster.

Since our method directly uses 2D Gaussians to represent scene surfaces, allowing the Gaussian splat to better adhere to the surface geometry, it outperforms 3DGS-based methods [9, 51]. Furthermore, while 2DGS [10] and some other methods [52, 53] that employ depth strategies do improve geometric reconstruction quality, they still struggle in indoor scenes due to the complexity of spatial structures and the prevalence of textureless regions. By integrating seed-guided strategies and geometric constraints, our method enhances the accuracy of scene structure capture and achieves higher reconstruction quality, resulting in superior metrics.

As shown in Fig. 4, some methods [4, 51] produce noisy reconstructions with scattered floaters, and fail to represent the actual surfaces accurately due to the lack of geometric constraints. However, they may cover more ground truth data and thus achieve higher Accuracy than 2DGS on the ScanNet++ dataset in Table 1. Our method improves the structural coherence of the reconstruction, leading to a more accurate representation of the scene and a significant improvement in the Accuracy metric compared to 2DGS.

5.3. Ablation Studies

To assess the individual contributions of each component in our model, we perform ablation studies on the ScanNet dataset. The quantitative results are reported in Table 2 and Figure 5 shows the qualitative results. These allow us to isolate the impact of key elements on the overall reconstruction quality.

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**

Table 2. Results of the ablation study on ScanNet dataset. The best results are marked in **bold**.

Seed Points Guidance. Figure 5 shows that without seed points guidance, the scene lacks clear structural organization, leading to a significantly inflated and disorganized reconstruction. Adding this module enables our method to better capture the underlying geometric framework of indoor scenes, improving the F-score by 87.3% in Table 2.

Monocular Depth Supervision. As shown in Figure 5, removing depth supervision leads to spatial misalignments and unrealistic arrangements. Incorporating depth supervision significantly enhances geometric accuracy, achieving a 31.3% F-score increase as reported in Table 2.

Monocular Normal Supervision. Removing normal supervision results in surface inconsistencies as shown in Figure 5, with certain planar areas like walls, floors, and doors misaligned. Adding this module improves surface alignment, increasing the F-score by 10.6% in Table 2.

Multi-View Consistency Constraints. Figure 5 reveals some Gaussians fail to align with the correct areas with the absence of multi-view constraints. Introducing this component reduces view-dependent inconsistencies to a certain degree, further enhancing the reconstruction quality.
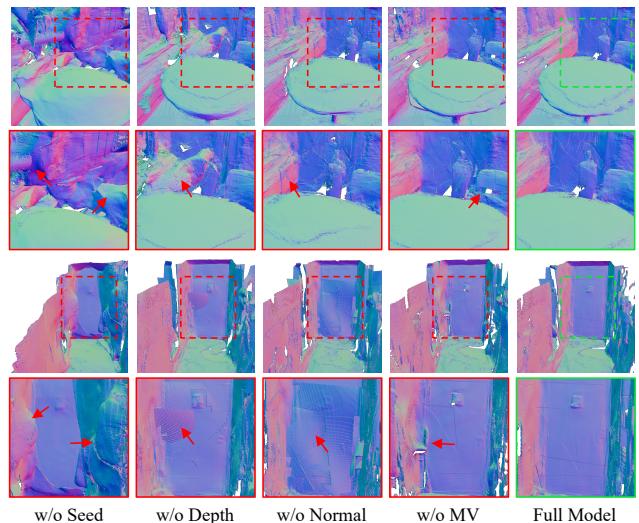

w/o Seed   w/o Depth   w/o Normal   w/o MV   Full Model

Figure 5. **Qualitative results of ablation study.**

6. Conclusion

We propose 2DGS-Room, a novel method for indoor scene reconstruction based on 2D Gaussian splatting by incorporating structural information from the scene to generate seed points, which guide the local Gaussian distributions. By leveraging geometric priors, we enhance the reconstruction quality of textureless regions and fine details in complex indoor environments. We also utilize multi-view consistency to reduce view-dependent inconsistencies to a certain degree. Extensive experiments show our method achieves superior performance compared with existing methods on multiple metrics and various indoor scenes.

References

[1] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 1, 2

[2] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5695–5703, 2016. 2

[3] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 1, 2

[4] 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. *NeurIPS*, 2021. 2, 6, 7, 8

[5] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. In *European Conference on Computer Vision*, pages 139–155. Springer, 2022. 2, 6

[6] 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. 2, 6

[7] Xinghui Li, Yikang Ding, Jia Guo, Xiansong Lai, Shihao Ren, Wensen Feng, and Long Zeng. Edge-aware neural implicit surface reconstruction. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1643–1648. IEEE, 2023.

[8] Xinghui Li, Yuchen Ji, Xiansong Lai, Wanting Zhang, and Long Zeng. Fine-detailed neural indoor scene reconstruction using multi-level importance sampling and multi-view consistency. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 3477–3483. IEEE, 2024. 2

[9] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):1–14, 2023. 2, 3, 4, 6, 7, 8

[10] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. *arXiv preprint arXiv:2403.17888*, 2024. 2, 3, 4, 6, 7, 8

[11] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 2

[12] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Gipuma: Massively parallel multi-view stereo reconstruction. *Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e. V*, 25 (361-369):2, 2016.

[13] Robust Multiview Stereopsis. Accurate, dense, and robust multiview stereopsis. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 32(8), 2010. 2

[14] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 2

[15] Adrian Broadhurst, Tom W Drummond, and Roberto Cipolla. A probabilistic framework for space carving. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, pages 388–393. IEEE, 2001. 2

[16] Jeremy S De Bonet and Paul Viola. Poxels: Probabilistic voxelized volume reconstruction. In *Proceedings of International Conference on Computer Vision (ICCV)*, page 2. Citeseer, 1999.

[17] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *International journal of computer vision*, 35:151–173, 1999.

[18] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2019–2028, 2020. 2

[19] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5038–5047, 2017. 2

[20] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2015.

[21] Gernot Riegler, Ali Osman Ulusoy, Horst Bischof, and Andreas Geiger. Octnetfusion: Learning depth fusion from data. In *2017 International Conference on 3D Vision (3DV)*, pages 57–66. IEEE, 2017.

[22] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2821–2830, 2018.

[23] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5525–5534, 2019.

[24] Zehao Yu and Shenghua Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1949–1958, 2020.

[25] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *arXiv preprint arXiv:2008.07928*, 2020. 2

[26] 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. 2

[27] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2

[28] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Pointnerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5438–5448, 2022.

[29] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023. 2

[30] 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. 2

[31] 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.

[32] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022.

[33] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022.

[34] 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. 2, 6, 7, 8

[35] 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. 2

[36] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5610–5619, 2021. 2

[37] 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. 2

[38] 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. 2

[39] Yixing Lao, Xiaogang Xu, Xihui Liu, Hengshuang Zhao, et al. Corresnerf: Image correspondence priors for neural radiance fields. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[40] 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. 2

[41] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 2

[42] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 2

[43] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 2

[44] 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. 2

[45] Jingyang Zhang, Yao Yao, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Critical regularizations for neural surface reconstruction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6270–6279, 2022. 2

[46] 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. 2

[47] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 3

[48] Matias Turkulainen, Xuqian Ren, Iaroslav Melekhov, Otto Seiskari, Esa Rahtu, and Juho Kannala. Dn-splatter: Depth and normal priors for gaussian splatting and meshing. *arXiv preprint arXiv:2403.17822*, 2024. 3

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

[50] 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. 3

10

[51] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *arXiv preprint arXiv:2311.12775*, 2023. 3, 6, 7, 8, 2

[52] 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. *arXiv preprint arXiv:2406.06521*, 2024. 3, 6, 7, 8

[53] Baowen Zhang, Chuan Fang, Rakesh Shrestha, Yixun Liang, Xiaoxiao Long, and Ping Tan. Rade-gs: Rasterizing depth in gaussian splatting. *arXiv preprint arXiv:2406.01467*, 2024. 3, 6, 7, 8, 2

[54] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 5

[55] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 5

[56] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13137–13146, 2021. 5

[57] 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. 6, 1

[58] 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. 6, 1

[59] Jae-Chern Yoo and Tae Hee Han. Fast normalized cross-correlation. *Circuits, systems and signal processing*, 28:819–843, 2009. 6

[60] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 6

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

Supplementary Material

In this supplementary material, we provide the following components:

- Definitions of the 3D geometry metrics used to evaluate reconstruction quality in Sec. A.
- Additional details of the datasets, training configuration, and the iteration schedule for key modules in Sec. B.
- Additional qualitative results, including mesh comparison, ablation results, and rendering comparison in Sec. C.

A. Definitions of Eevaluation Metrics

We evaluate our method using five widely-used 3D geometry metrics: Accuracy, Completion, Precision, Recall, and F-score, defined in Table 3. These metrics collectively assess the geometric fidelity of the reconstructed point clouds by measuring the alignment between the predicted and ground truth point clouds.

Accuracy measures the average distance between reconstructed points and the ground truth, with smaller values indicating better alignment. Completion assesses how well the reconstruction covers the ground truth, where lower values are better. Precision and Recall evaluate the proportion of points within a set threshold, with higher values indicating better performance. F-score, the harmonic mean of Precision and Recall, provides a balanced measure of reconstruction quality, where higher values reflect superior results.

Metric	Definition
Acc.	$\mathrm{mean}_{c \in C}(\min_{c^* \in C^*} \|c - c^*\|)$
Comp.	$\mathrm{mean}_{c^* \in C^*}(\min_{c \in C} \|c - c^*\|)$
Prec.	$\mathrm{mean}_{c \in C}(\min_{c^* \in C^*} \|c - c^*\| < .05)$
Recall	$\mathrm{mean}_{c^* \in C^*}(\min_{c \in C} \|c - c^*\| < .05)$
zoF-score	$\frac{2 \times \mathrm{Prec} \times \mathrm{Recall}}{\mathrm{Prec} + \mathrm{Recall}}$

Table 3. **Definitions of 3D metrics.** c and c^* are the predicted and ground truth point clouds.

B. Additional Implementation Details

Datasets. As described in the main paper, the quantitative evaluation metrics are derived from results tested two datasets. Specifically, we select 8 scenes from the ScanNet dataset [57]: scene0050_00, scene0085_00, scene0114_02, scene0580_00, scene0603_00, scene0616_00, scene0617_00, scene0721_00, and 4 scenes from the ScanNet++ dataset [58]: 8b5caf3398, 8d563fc2cc, 41b00feddb, b20a261fdf.

Training details. For all scenes, our seed-guided optimization is performed between 1,500 and 15,000 iterations. We set $N_g = 100$ for the gradient-guided growth and $N_\alpha = 100$ for the pruning strategy. Depth supervision and normal supervision are applied consistently from the first iteration through to the end of training, providing continuous geometric constraints. The multi-view consistency constraint is introduced after 7,000 iterations, once the foundational structure has been established, to further improve view alignment.

C. Additional Qualitative Results

C.1. Additional Ablation Results

To complement the local detail comparisons in the main paper, we provide additional ablation results focusing on the overall scene structure in Figure 6. These visualizations highlight the contributions of key components, including the seed points guidance, monocular depth supervision, and monocular normal supervision. The multi-view consistency constraints are primarily designed to further mitigate floating artifacts in certain scenarios, which have a limited impact on the overall structure. Therefore, they are not included in these structural comparisons. Their effectiveness is instead reflected in the qualitative results shown in Figure 5 and the quantitative metrics presented in Table 2 of the main paper.

When the seed points guidance strategy is removed, the reconstructed objects appear fused together, with unclear boundaries, compromising the scene's structural clarity.

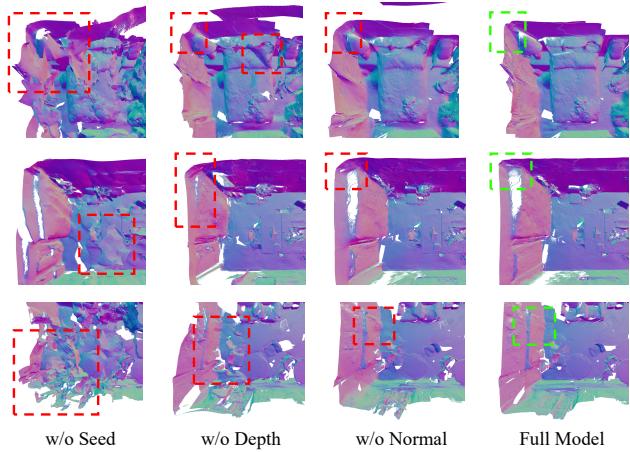w/o Seed   w/o Depth   w/o Normal   Full Model

Figure 6. **Additional qualitative results of ablation study.**

Without depth supervision, objects exhibit depth misalignments, leading to unrealistic spatial arrangements. Similarly, excluding normal supervision results in uneven surfaces, especially on planar regions like walls, where visible curvature or misalignment artifacts occur.

C.2. Additional Qualitative Comparison

In addition to the four indoor scenes shown in the main paper, we further include qualitative reconstruction comparison results of the different methods [4, 10, 51–53] on additional scenes from ScanNet and ScanNet++. As demonstrated in Figure 7, our method significantly outperforms other approaches in capturing global structures, preserving fine-grained details as well as reducing artifacts in textureless regions.

C.3. Rendering Comparison

We also provide extensive rendering results comparing our 2DGS-Room with 2DGS across various scenes and viewpoints from the ScanNet and ScanNet++ datasets in Figures 8, 9, and 10. Rendered RGB, depth, and normal maps are shown for visual comparison. Our method achieves significant improvements in the rendering quality of depth and normal maps, showcasing smoother transitions and more accurate surface details. Furthermore, the quality of the RGB images rendered by our method remains robust and shows clear advantages over 2DGS in challenging scenarios, such as handling fine details and varying lighting conditions. This demonstrates the effectiveness of our method in achieving superior geometric reconstructions while maintaining photometric accuracy.

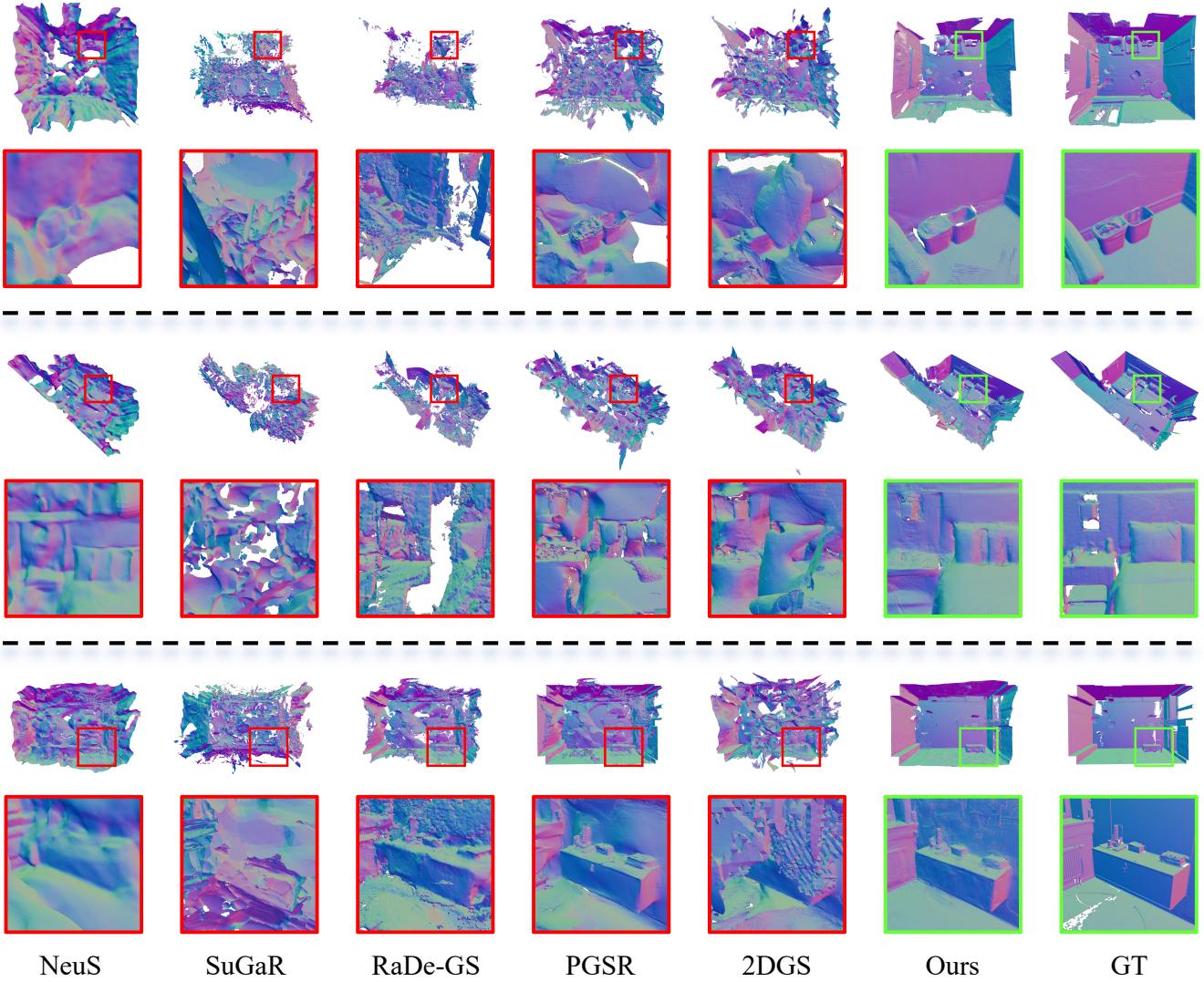| NeuS | SuGaR | RaDe-GS | PGSR | 2DGS | Ours | GT |

Figure 7. **Additional qualitative reconstruction comparison.** For each indoor scene, the first row is the top view of the whole room and the second row is the details of the masked region.
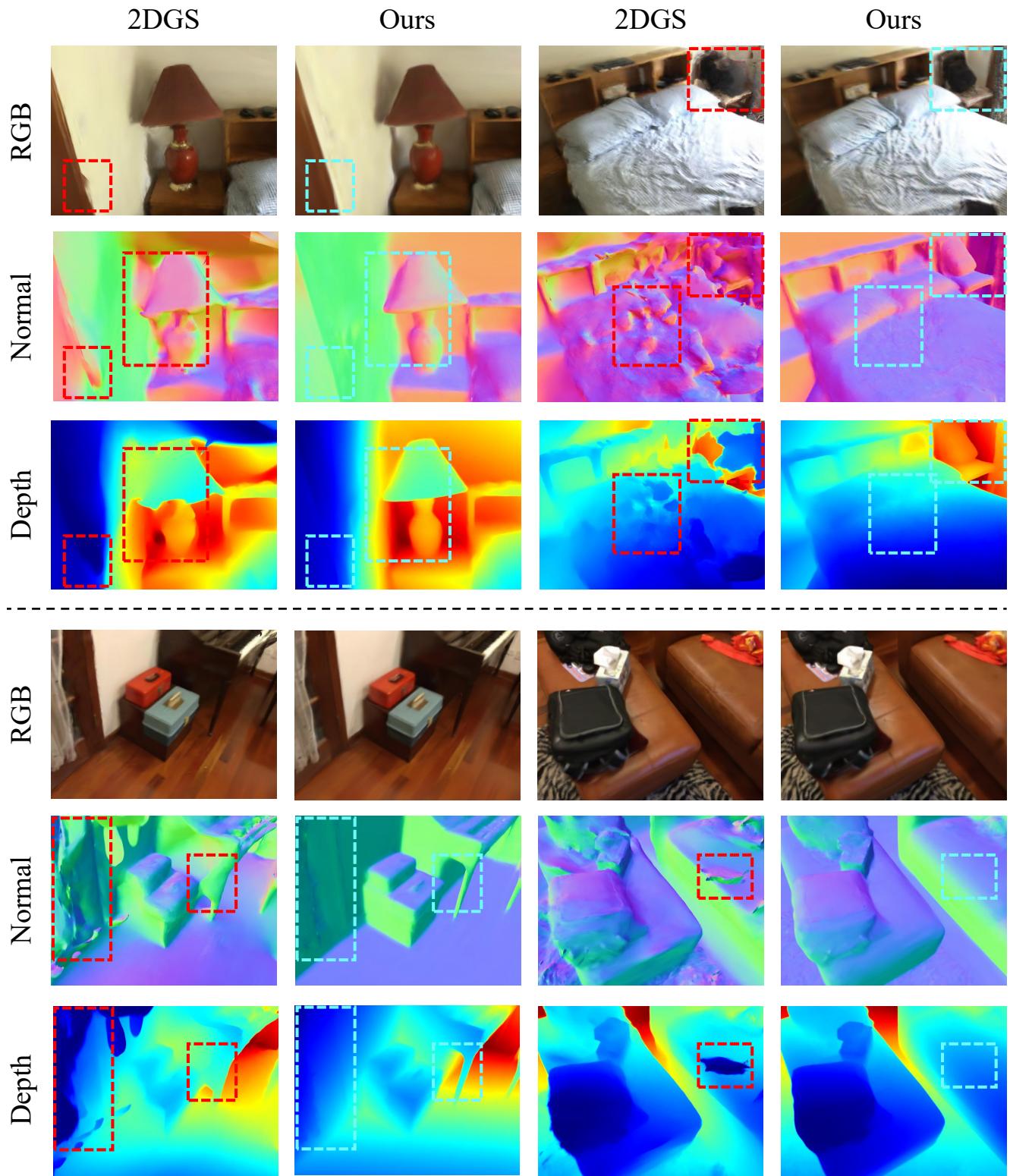
Figure 8. **Rendering comparison on the ScanNet dataset (scene0580 and scene0050).**
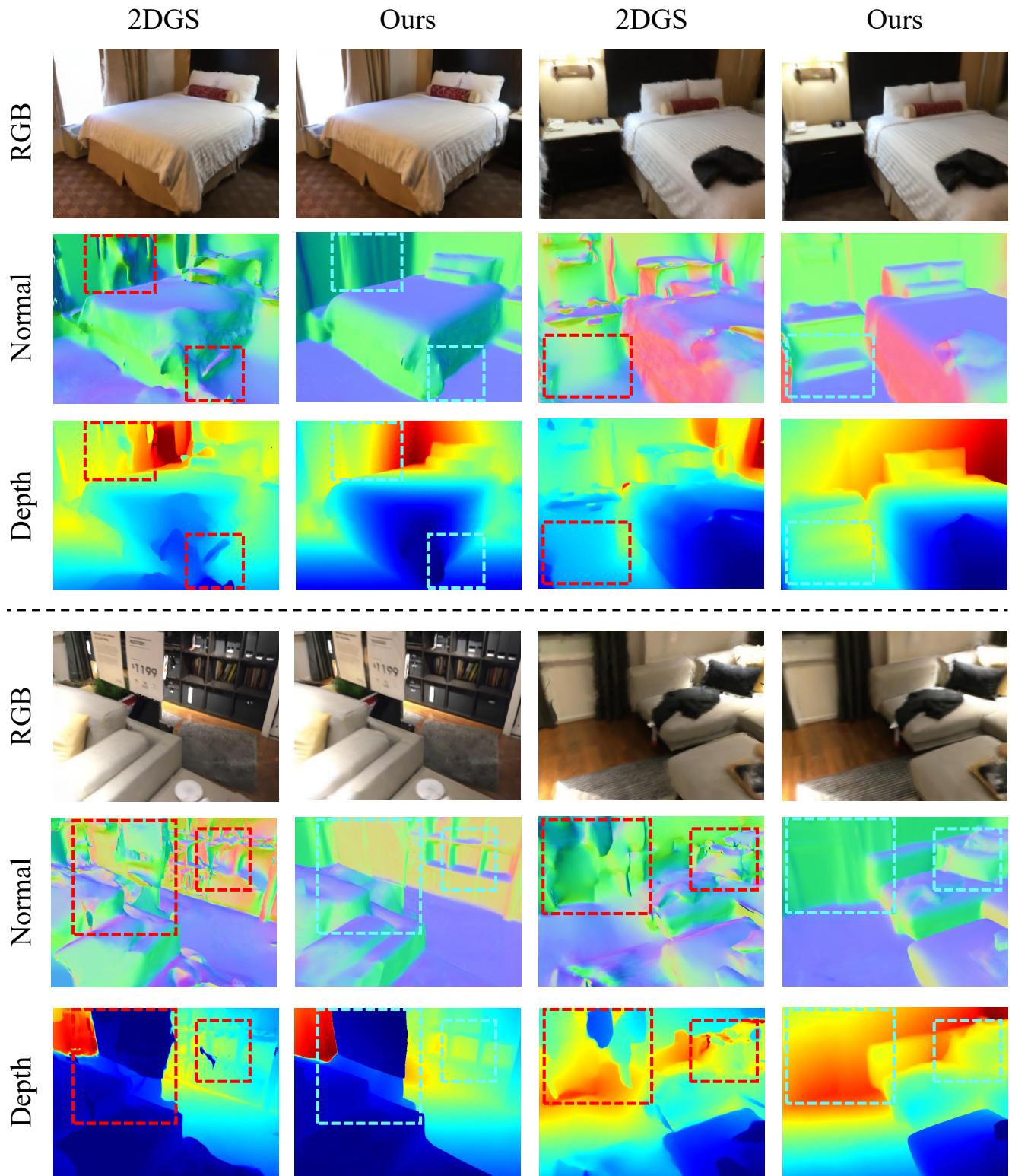
Figure 9. **Rendering comparison on the ScanNet dataset (scene0085 and scene0617).**
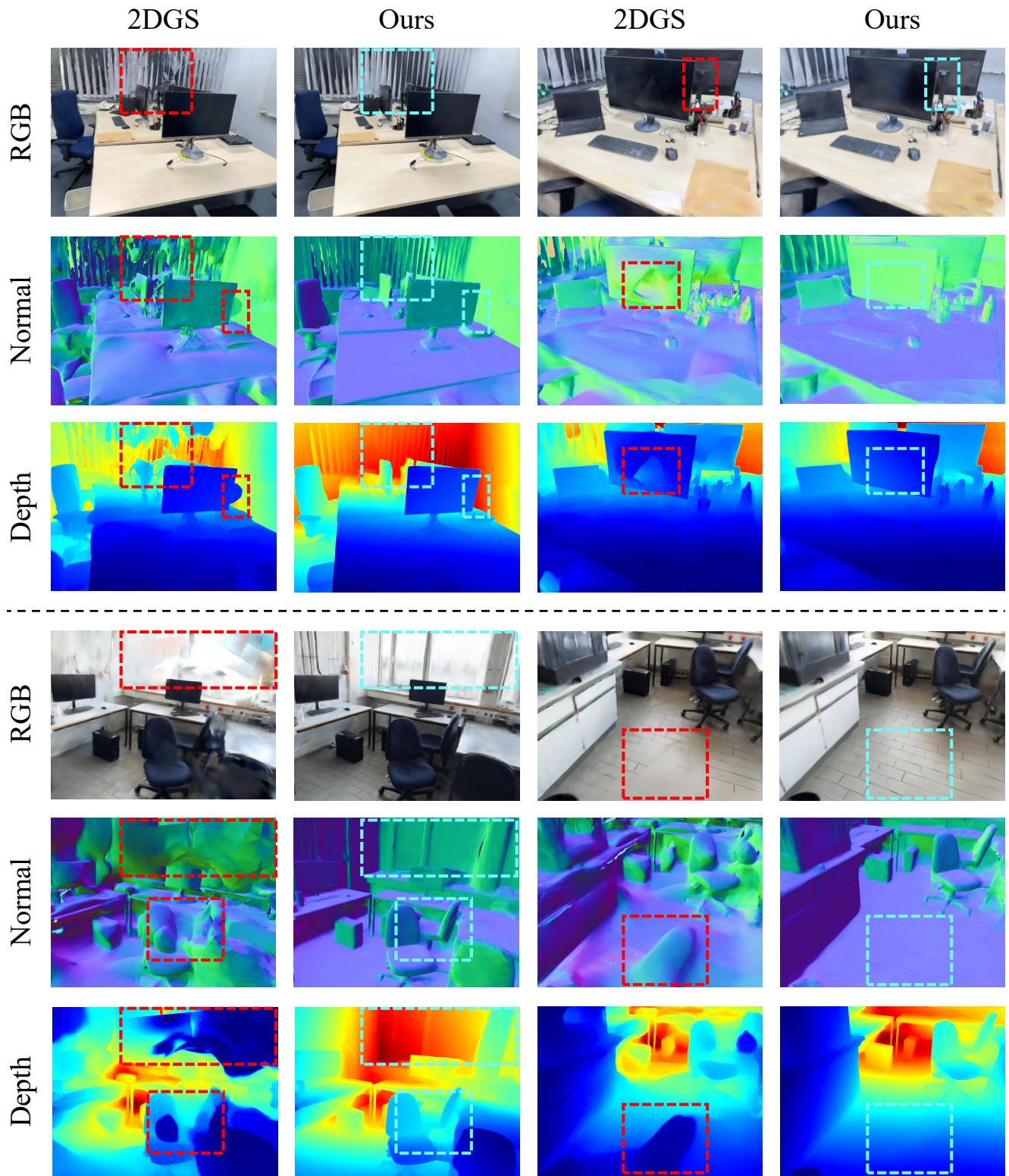
	2DGS	Ours	2DGS	Ours

Figure 10. **Rendering comparison on the ScanNet++ dataset (8d563fc2cc and 41b00feddb).**