IndoorSplat: Enhanced Indoor Scene Reconstruction with Structured 2D Gaussian Splatting

Min Shi[1], Xigang Zhao[1⋆], Tao Yang[1⋆], Qi Wang[1✉], Dengming Zhu[2,3], and Zhaoxin Li[4]

[1] School of Control and Computer Engineering, North China Electric Power University, China
[2] Institute of Computing Technology, Chinese Academy of Sciences, China
[3] Taicang-ZK Institute of Information and Technology, China
[4] Agricultural Information Institute of CAAS, Beijing, China

Abstract. Geometric reconstruction of indoor scenes poses unique challenges due to their inherent characteristics, such as large textureless surfaces, complex occlusions, and intricate geometric structures. While recent 2D/3D Gaussian Splatting-based methods have achieved impressive performance in general scene reconstruction, their direct application to indoor settings often leads to noisy, incomplete, or geometrically inconsistent results. In this paper, we propose a novel framework **IndoorSplat** to bridge this gap and enable high-quality indoor scene geometry reconstruction based on 2D Gaussian Splatting (2DGS). To this end, we first revisit the importance of Gaussian initialization in the indoor scene reconstruction process and introduce a carefully designed dense initialization strategy. This strategy leverages local geometric priors to densify point clouds estimated by an advanced feed-forward model and utilizes them to initialize the structure of Gaussian primitives. In addition, to better capture the rich structural details in indoor scenes, we introduce an iterative adaptive densification strategy that accelerates convergence in complex regions by applying more aggressive splitting and cloning policies to Gaussians with consistent activation. Extensive experiments demonstrate that our method achieves state-of-the-art performance by producing more geometrically accurate and complete surface reconstructions compared to existing approaches.

Keywords: Gaussian Splatting · Indoor Scene Reconstruction · Surface Reconstruction

1 Introduction

High-fidelity surface reconstruction of indoor environments remains a fundamental yet inherently challenging task in computer vision and computer graphics,

⋆Equal contribution.
✉Corresponding author.

with broad applications in virtual reality [10], robotics [1], and digital heritage preservation [8]. Recent advances in 3D Gaussian Splatting (3DGS) [17] have demonstrated impressive rendering efficiency and photorealistic quality in novel view synthesis, particularly for richly textured outdoor scenes. However, when applied to indoor scene reconstruction, its performance degrades noticeably. Indoor environments often exhibit large textureless surfaces, intricate geometry, and fine-grained details, which significantly increase the ambiguity between geometry and appearance. As a result, 3DGS tends to produce incorrect geometry-appearance combinations and floating artifacts that satisfy photometric loss but lead to incomplete and noisy surface reconstructions.

While several extensions of 3D Gaussian Splatting, *e.g.*, 2DGS [37], GOF [35], PGSR [4], have been proposed to improve geometric reconstruction, their performance remains suboptimal in indoor environments. Overall, three main factors contribute to this limitation. First, these methods are heavily reliant on an initial point cloud generated by conventional Structure-from-Motion (SfM) methods [22]. However, in textureless indoor regions, SfM often struggles to find reliable correspondences, leading to sparse, noisy, or missing points that impair the initialization of Gaussian primitives and subsequent optimization. Second, the default densification and pruning strategies, particularly the global opacity reset, can be overly naive. While effective at suppressing floaters, they fail to facilitate convergence in fine-grained geometric regions and may inadvertently remove valid geometric structures. Third, the exclusive reliance on photometric loss fails to disambiguate geometry and appearance, often resulting in suboptimal local minima for both. The absence of strong geometric constraints further reduces reconstruction robustness in complex indoor settings.

To overcome these core limitations, we propose a novel framework **Indoor-Splat** that extends 2DGS for high-quality indoor surface reconstruction. The method begins with a dense initialization strategy, where an advanced feed-forward model [9] provides a strong structural prior to generate a more complete point cloud, surpassing the limitations of traditional SfM. However, in textureless or sparsely observed regions, the feed-forward model may still fail to produce reliable geometry. To address this, a novel densification strategy is proposed. The initial points are projected into multiple views to identify sparse image patches, and additional 3D points are then back-projected from locally aligned monocular depth priors [3], effectively densifying these regions and filling geometric holes. The resulting dense structural point cloud is then used to carefully initialize the properties of the Gaussian primitives, including their positions, normals, scales, and colors, providing a geometrically sound foundation for subsequent optimization.

To better capture fine geometric details, a novel iterative adaptive densification strategy is introduced during training. Newly split or cloned Gaussian primitives are tracked and marked as active. These active primitives naturally tend to fall in geometrically intricate regions and are thus prioritized for further refinement. This focused densification enhances detail preservation while avoiding unnecessary redundancy. To stabilize the process, the disruptive global opacity reset used in prior methods is replaced with a soft opacity regularization, which gradually suppresses redundant Gaussian primitives without harming well-formed structures. To further enhance the accuracy and completeness of reconstruction, additional geometric supervision is introduced by leveraging monocular depth and normal priors.

Our main contributions can be summarized as follows:

- A novel framework tailored for indoor scene reconstruction, achieving state-of-the-art (SOTA) performance on challenging indoor datasets, yielding surface reconstructions with significantly improved geometric accuracy and completeness compared to prior methods.
- A dense and structured Gaussian primitives initialization strategy that combines a feed-forward model with geometric priors. This design highlights the critical role of high-quality initialization for accurate indoor surface reconstruction.
- An iterative adaptive densification strategy, combined with a stable opacity regularization, that enhances the representation of fine-scale scene details and improves overall reconstruction robustness.

2 Related Work

2.1 Surface Reconstruction

To achieve surface reconstruction, neural implicit surfaces-based methods [24, 25, 29, 34] have introduced signed distance functions (SDF). However, these approaches often suffer from slow training speeds due to the need for extensive ray sampling, resulting in high computational costs. Recently, 3D Gaussian plating (3DGS) [17] has emerged as a popular alternative, offering high efficiency in both training and rendering. 3DGS represents the scene as a set of Gaussian primitives and utilizes a splatting algorithm for fast and differentiable rendering, achieving impressive performance in novel view synthesis.

While many efforts [5, 28, 33, 38] have focused on enhancing the rendering quality of 3DGS, its explicit representation of geometry also presents a strong foundation for surface extraction. Consequently, another research direction has focused on extending 3DGS for surface reconstruction. These approaches typically enhance 3DGS to compute depth and normal information, which is then used in mesh extraction algorithms such as Truncated Signed Distance Function (TSDF) fusion [19] or Poisson Surface Reconstruction [15,16]. Methods like SuGaR [12], High-fidelity Gaussian Splatting [7], and PGSR [4] regularize the scale of Gaussians, making depth and normal computation more tractable. In contrast, 2D Gaussian Splatting (2DGS) [14] directly replaces standard Gaussian primitives with 2D-Gaussian representations, allowing for a natural and accurate computation of depth and normals without additional regularization, which significantly improves reconstruction quality.

2.2 Indoor Surface Reconstruction

However, the methods described above often fail in indoor scenes due to the prevalence of large textureless regions. In such environments, relying solely on photometric loss from standard 3DGS optimization fails to capture accurate geometry, resulting in degraded reconstruction quality. To address this issue, some recent approaches incorporate monocular depth and surface normal priors to improve reconstruction performance. DN-Splatter [23] integrates sensor depth and monocular depth estimates to apply stronger geometric constraints in texture-poor regions. Additionally, it uses normals from monocular networks to guide Gaussian alignment, improving geometric accuracy in challenging indoor scenes. IndoorGS [21] leverages geometric information embedded in monocular priors to densify the initial point cloud and assigns different Adaptive Densification Control (ADC) strategies to Gaussians based on semantic categories. Gaussian-Room [27], on the other hand, extracts normal and edge priors from a single image to provide geometric cues in textureless planar regions and high-frequency detail areas by introducing additional loss constraints and guiding the training sample selection process. AGS-Mesh [20] fuses low-resolution sensor depth with monocular normal priors from smartphones, using a Depth-Normal Consistency filter to remove unreliable depth data by cross-checking sensor-derived normals against monocular predictions. This adaptive regularization handles noise and improves reconstruction.

Despite these advancements, most existing methods still suffer from poor initialization, as they typically rely on sparse point clouds generated by traditional Structure-from-Motion (SfM) [22] pipelines. Inadequate initialization makes it difficult for the model to reconstruct accurate surfaces.

3 Preliminaries

3.1 2D Gaussian Splatting

2D Gaussian Splatting represents a 3D scene using a set of 2D oriented planar Gaussians. Each 2DGS primitive can be considered a local tangent plane $P(u, v)$. It is parameterized by a center point $\mathbf{p}_k$, two principal tangential vectors $\mathbf{t}_u$ and $\mathbf{t}_v$, and a scaling vector $\mathbf{S} = (s_u, s_v)$ that controls the variances of the 2D Gaussian:

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

For a point $\hat{\mathbf{u}} = (u, v)$ in the local uv-space of a 2DGS, its value is calculated using a standard Gaussian function:

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

Its splatting process can be described by a 2D-to-2D transformation in homogeneous coordinates, where $\mathbf{x}$ represents the position in the pixel coordinate system, $\mathbf{W} \in 4 \times 4$ is the combined transformation matrix from world space to screen space:

$$\mathbf{x} = \mathbf{W}P(u, v). \tag{3}$$

To avoid computational instability, 2DGS transforms each pixel's viewing ray into the local 2D coordinate system of the Gaussian primitive. This is achieved by parameterizing the ray as the intersection of two orthogonal planes, which enables a direct, analytical solution for the ray-Gaussian intersection point. The rendering process for 2DGS is similar to that of 3DGS. With o_i representing the opacity of the Gaussian, the alpha value is computed as:

$$\alpha_i = o_i \mathcal{G}(\hat{\mathbf{u}}). \tag{4}$$

Then, the color for each pixel $\mathbf{x}$ is computed via volumetric alpha-blending from front to back:

$$\hat{C}(\mathbf{x}) = \sum_{i \in N} c_i \alpha_i T_i, \quad \text{where } T_i = \prod_{j=1}^{i-1} (1 - \alpha_j). \tag{5}$$

Here, T_i is the accumulated transmittance at the pixel location, $\mathbf{c}_i$ is a view-dependent color represented by Spherical Harmonics (SH) coefficients, and N denotes the number of Gaussian primitives.

3.2 Rasterizing Depth and Normal in 2DGS

For a point $\mathbf{x}$ in uv-space, the normal $\hat{\mathbf{N}}(\mathbf{x})$ is calculated using alpha-blending, which is a process similar to color rendering:

$$\hat{\mathbf{N}}(\mathbf{x}) = \sum_{i \in N} \mathbf{n}_i \alpha_i T_i, \tag{6}$$

where $\mathbf{n}_i$ is the normal vector of the i-th Gaussian. The expected depth $\hat{\mathbf{D}}(\mathbf{x})$ is computed from the depth samples d_i and their corresponding weights ω_i:

$$\hat{\mathbf{D}}(\mathbf{x}) = \frac{\sum_i \omega_i d_i}{\sum_i \omega_i + \epsilon}, \tag{7}$$

where $\omega_i = T_i \alpha_i$ is the weight contribution of the i-th Gaussian, d_i represents the depth of the i-th Gaussian, and ϵ is a small positive constant used to prevent division by zero.

4 Method

Our primary objective is to enhance 2D Gaussian Splatting for the high-fidelity surface reconstruction of complex indoor scenes. To this end, we first propose a dense initialization strategy (Sec. 4.1). Next, we introduce a novel optimization strategy featuring two core components (Sec. 4.2): an iterative adaptive densification strategy and a stable L_1 opacity regularization. Finally, the entire optimization process is guided by explicit geometric prior losses derived from monocular depth and normal estimates (Sec. 4.3). An overview of our framework is provided in Fig. 1.

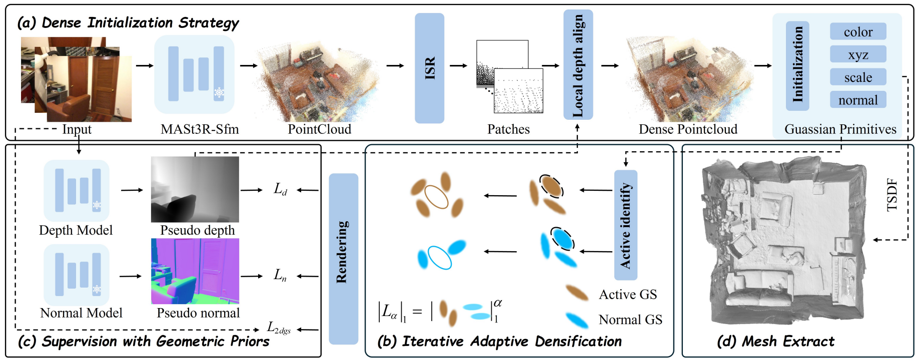

Fig. 1: An overview of IndoorSplat pipeline. (a) We generate an initial point cloud using MASt3R-SfM. The point cloud is then projected to 2D, and sparse regions are identified using the Identify Sparse Region (ISR) module. Finally, local depth alignment is applied to densify the point cloud. (b) The initial Gaussians are refined through iterative adaptive densification and a stable L_1 opacity regularization for non-destructive pruning. (c) The entire optimization is guided by both the photometric loss and explicit geometric prior losses derived from monocular depth and normals.

4.1 Dense Initialization Strategy

Since traditional SfM methods often generate overly sparse point clouds (as shown in Fig. 2 (a)) that result in a poor initialization of Gaussian primitives, we begin with a dense initialization strategy to enhance the point cloud for better capturing the scene's geometric structure. First, we leverage MASt3R-SfM [9], an advanced feed-forward model, to obtain initial point clouds $\mathbf{P}_{\text{sfm}}$ and the corresponding camera parameters, including the intrinsic matrix $\mathbf{K}$, rotation $\mathbf{R}$, and translation $\mathbf{T}$. For brevity, we omit subscripts throughout the method section. In contrast to traditional SfM methods [22], MASt3R-SfM directly generates point clouds followed by global alignment, offering improved robustness in indoor environments with weak textures.

However, MASt3R-SfM still exhibits limitations in weakly observed regions, often resulting in holes within the point cloud (see Fig. 2 (b)). To overcome this issue, we subsequently propose a dense initialization strategy, which leverages depth priors to locally align and densify sparse regions, thereby improving the overall quality of the initial point cloud. The entire process consists of two steps: first, identifying the sparse regions that require completion, and second, generating dense point clouds for these areas.

To accurately identify regions that require densification, we first project the sparse point cloud $\mathbf{P}_{\text{sfm}}$ onto all input images $\mathbf{I}$ and calculate their depth $\mathbf{D}_{\text{sfm}}$ corresponding to each input image:

$$d \begin{bmatrix} \mathbf{u} \\ 1 \end{bmatrix} = \mathbf{K} \left(\mathbf{R} \mathbf{P}_{\text{sfm}} + \mathbf{T} \right), \tag{8}$$

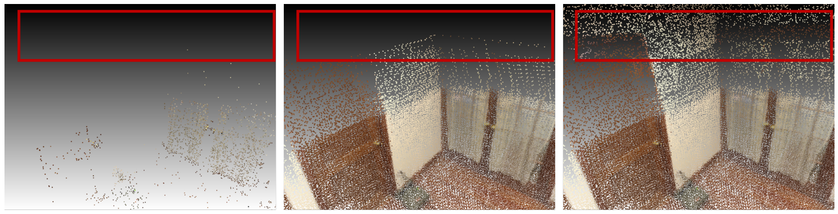

Fig. 2: Comparison of point clouds. (a) COLMAP produces a very sparse reconstruction. (b) MASt3R-SfM recovers the basic geometric structure. (c) Our dense initialization strategy generates an accurate and detailed point cloud.

where $\mathbf{u}$ and $d \in \mathbf{D}_{\text{sfm}}$ are the 2D coordinates and the corresponding depth of $\mathbf{P}_{\text{sfm}}$. To effectively handle occlusions, we do not retain all projected points; instead, we perform point culling relative to the camera pose, explicitly discarding points that fall outside the field of view or are occluded by other objects. Next, we partition each input $\mathbf{I}$ into $N \times N$ patches $\mathbf{C}_{i,j}$ and count the number of projected points in each patch. **Patches with a point count below a predefined threshold τ are marked as candidates for densification.** To ensure reliability, we exclude patches that contain sharp depth discontinuities in the monocular depth $\mathbf{D}$ obtained from a pretrained depth estimation network, as the scale ambiguity between foreground and background in such regions leads to unreliable affine transformations.

For each sparse patch $\mathbf{C}_{i,j}$, we perform densification guided by the monocular depth $\mathbf{D}$. Since monocular depth estimation often suffers from scale inconsistency, we introduce a robust local alignment algorithm to mitigate this issue. As a sparse patch $\mathbf{C}_{i,j}$ contains too few projected SfM points for stable alignment, we first expand the patch $\mathbf{C}_{i,j}$ outward by half the patch size in all directions at each step until a sufficient number of SfM points are included in the region, resulting in an extended patch $\mathbf{C'}_{i,j}$. Then we solve a local least-squares problem to estimate the optimal rigid transformation $(\hat{s}, \hat{\mathbf{t}})$ using the SfM depth $\mathbf{D}_{\text{sfm}}$:

$$(\hat{s}, \hat{\mathbf{t}}) = \arg\min_{s,\mathbf{t}} \sum_{\mathbf{p} \in \mathbf{C'}_{i,j}} \left\| (s \cdot \mathbf{D}(\mathbf{p}) + \mathbf{t}) - \mathbf{D}_{\text{sfm}}(\mathbf{p}) \right\|_2^2 . \tag{9}$$

Next, we uniformly sample 2D pixel locations $\mathbf{u'}$ within the original patch $\mathbf{C}_{i,j}$ and compute their depth as $d = \hat{s} \cdot \mathbf{D}(\mathbf{u'}) + \hat{\mathbf{t}}$. These 2D points are finally back-projected to 3D space and added to the SfM point cloud to complete the region:

$$\mathbf{P}_{\text{add}} = \mathbf{R}^T \left(d \cdot \mathbf{K}^{-1} \begin{bmatrix} \mathbf{u'} \\ 1 \end{bmatrix} - \mathbf{T} \right), \tag{10}$$

$$\mathbf{P}_{\text{sfm}} \leftarrow \mathbf{P}_{\text{sfm}} \cup \mathbf{P}_{\text{add}}, \tag{11}$$

where $\mathbf{P}_{\text{add}}$ is the point cloud used to densify the region $\mathbf{C}_{i,j}$. The above process is repeated for every view to enhance the entire initial point cloud.

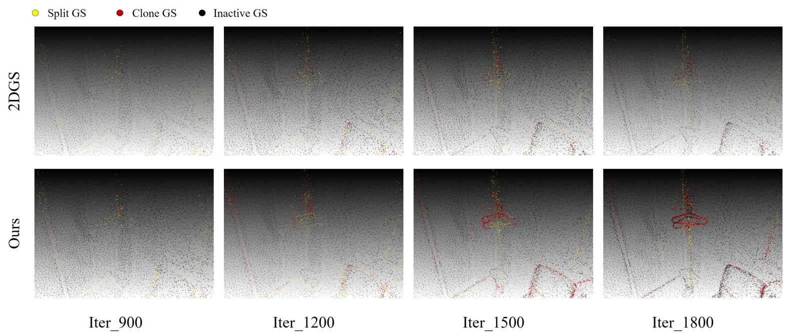

Fig. 3: Effectiveness of the ADC strategy.

After densifying the point cloud, we further enhance its geometric structure by initializing normals for the points. These normals are computed using Principal Component Analysis (PCA) [13]. To resolve the inherent directional ambiguity of PCA normals, we unify their orientation with respect to the point cloud's centroid $\mathbf{p}_c$. Specifically, for each point $\mathbf{p} \in \mathbf{P}_{\text{sfm}}$ we flip its normal $\mathbf{n}_k$ if $\mathbf{n}_k \cdot (\mathbf{p}_c - \mathbf{p}) < 0$, ensuring that most normals point towards the scene's interior.

Finally, we initialize not only the positions, colors, and scales of the Gaussian primitives, but also their orientations from the dense point cloud, ensuring more accurate normal directions from the very beginning of training.

4.2 Iterative Densification with Stable Opacity Control

The vanilla Gaussian Splatting approach treats all Gaussian primitives equally in its Adaptive Densification Control (ADC) strategy, applying a uniform densification policy across the entire scene. Furthermore, it employs a pruning strategy that periodically resets the opacity of all Gaussians to zero to remove floaters. This indiscriminate method leads to slow convergence in regions with complex geometry and causes well-reconstructed geometric details to be erroneously pruned as floaters. To address this, we propose a simple yet effective iterative densification strategy that identifies active Gaussians in detail-rich regions early in training and implements more aggressive splitting and cloning (see Fig. 3). Thanks to our high-quality initial point cloud, the Gaussians that require further densification consistently emerge in the same areas after a few iterations. These regions are typically geometrically intricate and require more Gaussian primitives for accurate surface approximation.

The core idea of our ADC strategy is to **leverage the principle of temporal locality**. To this end, we introduce an active state flag for each Gaussian primitive to track its status throughout optimization. Specifically, when a Gaussian primitive meets the densification criterion, it is either split or cloned depending on its scale. Both the original and newly generated Gaussian primitives are marked as active. In subsequent densification iterations, our algorithm prioritizes these active Gaussian primitives. If an active Gaussian primitive again meets the densification criterion, we apply a more aggressive proliferation policy, expanding it into 4 Gaussian primitives, whose positions are sampled from the original Gaussian's distribution (see Fig. 1b). This forms a positive feedback loop that strengthens the reconstruction of fine geometric details in complex regions. Conversely, stable Gaussian primitives that remain unchanged for a prolonged period revert to the default densification strategy. This strategy, driven by temporal locality, improves both convergence speed and the model's capacity to capture fine-grained geometry.

Moreover, Gaussian Splatting-based approaches periodically reset the opacity of all Gaussians. While effective in removing floaters, this aggressive strategy can inadvertently prune valid geometric structures, degrading visual quality. Inspired by [36], we adopt a global opacity regularization, defined as:

$$\mathcal{L}_\alpha = \lambda_\alpha \sum_i |\alpha_i|_1, \tag{12}$$

where α_i denotes the opacity of the i-th Gaussian, and λ_α controls the pruning strength. Unlike hard resets, our method acts as a soft constraint, penalizing opacity values directly. Gaussians that contribute little to rendering are naturally suppressed during training, while those essential for reconstruction retain their opacity. This results in better preservation of fine details and overall visual fidelity (see Fig. 5).

4.3 Geometric Priors Supervision

While photometric loss is highly effective for novel view synthesis, it only provides an indirect constraint on the geometry. This may cause the geometric structure to get trapped in a local minima, even when the rendering appears correct. Consequently, errors in the geometric structure often arise in low-texture and complex regions, which are common in indoor scenes. Therefore, we introduce geometric constraints from monocular depth and normal priors to further enhance the quality of the scene reconstruction.

Monocular Depth Supervision. We introduce monocular depth priors to constrain the geometry of the scene. However, due to the inherent scale and shift ambiguity between the rendered depth map $\hat{\mathbf{D}}$ and the reference monocular depth map $\mathbf{D}$, directly applying affine transformations can lead to depth distortions. To mitigate scale ambiguity, we employ the Pearson Correlation Coefficient (PCC) [2] to construct the depth loss, as PCC measures only the linear correlation between variables and is invariant to absolute scale and shift.

Our depth loss function consists of two components:

$$\begin{aligned} \mathcal{L}_d &= \sum \left(1 - \mathrm{PCC}(\mathbf{D}, \hat{\mathbf{D}})\right) + \lambda_s \cdot \mathcal{L}_s \\ \text{where } \mathcal{L}_s &= \frac{1}{HW} \sum_{i,j} (|\nabla_x \hat{\mathbf{D}}_{i,j}| + |\nabla_y \hat{\mathbf{D}}_{i,j}|). \end{aligned} \tag{13}$$

Here H and W are the height and width of the image, and ∇_x and ∇_y represent the horizontal and vertical image gradients. The PCC term ensures structural consistency between the rendered and monocular depths, preserving the geometric integrity. The smoothness regularization term $\mathcal{L}_s$ penalizes the gradients of the rendered depth $\hat{\mathbf{D}}$, encouraging smooth depth transitions across object surfaces.

Monocular Normal Supervision. Correct normal directions can significantly improve the quality of surface reconstruction, particularly in indoor scenes that contain numerous planar structures. Therefore, we introduce a monocular normal prior to supervise the orientation of the Gaussian primitive, thereby enhancing the realism and smoothness of the scene's surfaces. We use a cosine similarity loss to penalize the angular difference between the rendered normals $\hat{\mathbf{N}}$ and the monocular normals $\mathbf{N}$, which is predicted by a pre-trained normal estimation model. The normal supervision loss, $\mathcal{L}_\mathrm{n}$ is defined as follows:

$$\mathcal{L}_\mathrm{n} = \sum \left(1 - \frac{\hat{\mathbf{N}} \cdot \mathbf{N}}{\|\hat{\mathbf{N}}\| \cdot \|\mathbf{N}\|} \right). \tag{14}$$

4.4 Loss Function

Our total loss function L is a weighted sum of the following components:

$$\mathcal{L} = \mathcal{L}_{\mathrm{2dgs}} + \lambda_d \mathcal{L}_d + \lambda_n \mathcal{L}_n + \lambda_\alpha \mathcal{L}_\alpha, \tag{15}$$

where $\mathcal{L}_{\mathrm{2dgs}}$ is the standard loss term from the 2DGS paper, including a L_1 loss term and a D-SSIM loss term on the RGB image, along with a normal consistency regularization term. We set the weights for our geometric priors as $\lambda_d = 0.2$ for depth supervision (with an associated smoothness regularization weight of $\lambda_s = 1.0$) and $\lambda_n = 0.05$ for normal supervision. Additionally, the opacity regularization weight λ_α is set to follow a linear warm-up schedule, which gradually increases from 1×10^{-10} to 1×10^{-9} throughout the training process.

5 Experiments

5.1 Experimental Setup

Dataset. We evaluate the performance of our approach on both geometry reconstruction and rendering quality using two widely used public datasets, ScanNet [6] and ScanNet++ [31], selecting 6 real-world indoor scenes from each to form a total of 12 evaluation scenes.

Table 1: Performance comparison on the ScanNet and ScanNet++ datasets. The first , second and third best results in each column are highlighted. For DN-Splatter and AGS-Mesh, "–" indicates no valid results.

Method	ScanNet					ScanNet++				
	Acc↓	Comp↓	Prec↑	Recall↑	F1↑	Acc↓	Comp↓	Prec↑	Recall↑	F1↑
3DGS	0.1543	0.1731	0.3501	0.3666	0.3574	0.3497	0.1684	0.3150	0.4219	0.3579
+MASt3R	0.1431	0.0661	0.4315	0.6191	0.5066	0.1743	0.0628	0.4926	0.6801	0.5689
GOF	0.1185	0.1769	0.4352	0.4194	0.4250	0.0978	0.0838	0.5472	0.6515	0.5923
+MASt3R	0.1143	0.0835	0.4923	0.5985	0.5385	0.1294	0.0350	0.6395	0.8157	0.7155
SuGaR	0.1031	0.1307	0.4810	0.4292	0.4523	0.1965	0.0812	0.5512	0.6060	0.5752
+MASt3R	0.0808	0.0586	0.5782	0.6637	0.6155	0.0884	0.0488	0.6091	0.6886	0.6452
2DGS	0.1037	0.1458	0.4920	0.4190	0.4514	0.1227	0.1050	0.5869	0.6106	0.5978
+MASt3R	0.0943	0.0582	0.6263	0.6505	0.6368	0.0832	0.0376	0.7440	0.8198	0.7792
PGSR	0.0868	0.1348	0.5368	0.4538	0.4905	0.0947	0.0730	0.6100	0.6837	0.6428
+MASt3R	0.0969	0.0636	0.6060	0.6248	0.6140	0.0950	0.0353	0.7248	0.8358	0.7739
DN-Splatter	-	-	-	-	-	0.0488	0.0575	0.8453	0.8311	0.8357
AGS-Mesh	-	-	-	-	-	0.0731	0.0506	0.8193	0.8565	0.8346
Ours	0.0374	0.0420	0.7844	0.7352	0.7587	0.0319	0.0227	0.8735	0.9216	0.8967

Implementation Details. We train all models for 30k iterations. We set $\tau = 300$ for the dense initialization strategy. Geometric prior supervision is applied starting from the 7k iterations, utilizing monocular depth and normal priors generated by DepthPro [3] and StableNormal [30]. To extract the final surface, we render depth maps from all training viewpoints, fuse them using TSDF [19], and extract the mesh via the Marching Cubes algorithm [18]. All other hyperparameters remain consistent with the original 2DGS implementation. All experiments were conducted on a single machine equipped with an NVIDIA RTX 3090 GPU.

Metrics. Following existing methods [24, 27], we choose five metrics to evaluate the quality of the reconstructed surface: Accuracy (Acc), Completion (Comp), Precision (Prec), Recall, and F-score (F1) with a threshold of 5cm.

Baselines. We compare our method with five state-of-the-art Gaussian Splatting-based general surface reconstruction methods, including 3DGS [17], GOF [35], SuGaR [12], 2DGS [37], and PGSR [4]. Each method is evaluated using both COLMAP-based and MASt3R-based initialization. In addition, we compare to two RGBD-based methods specifically designed for indoor scene reconstruction on the ScanNet++ dataset, DN-Splatter [23] and AGS-Mesh [20]. Due to the lack of publicly available source code for IndoorGS [21], we are unable to report its performance for comparison.

5.2 Results and Evaluation

Quantitative Comparison. We first quantitatively compare our method with the baselines. As shown in Tab. 1, our approach achieves the best results across all

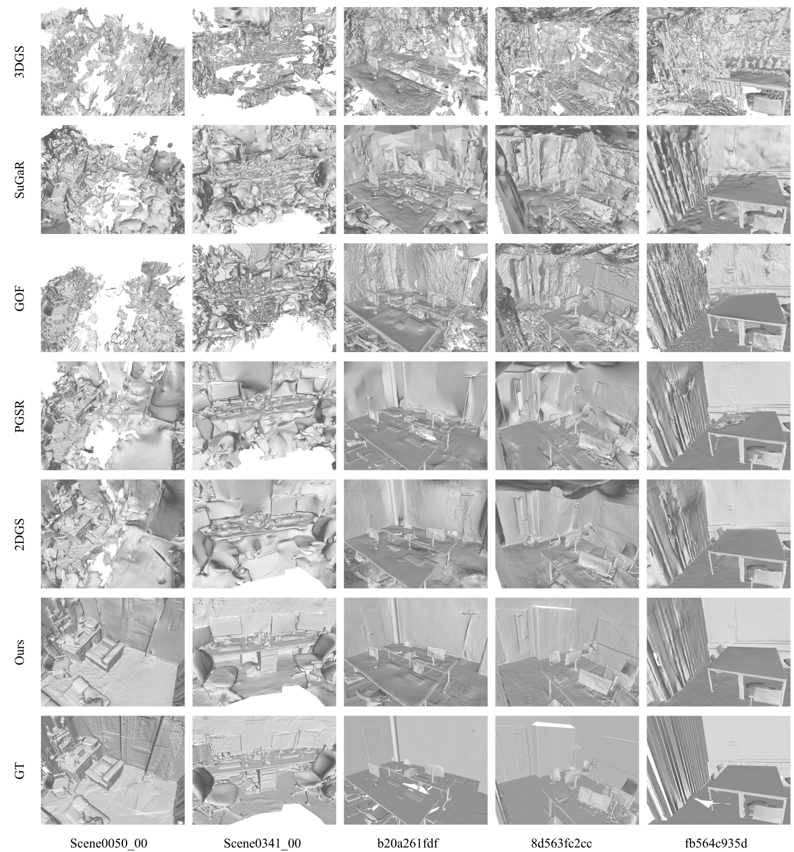

Fig. 4: Qualitative comparison of geometry on ScanNet (left two columns) and ScanNet++ (right three columns). Results demonstrate that our method achieves superior reconstruction performance in textureless regions and recovers more geometric details.

metrics on both datasets, demonstrating the effectiveness of our framework. The "+MASt3R" variants replace the Gaussian initialization from COLMAP with MASt3R, which consistently improves the baselines and highlights the strong influence of initialization quality on the final geometry. Despite this improvement, our method still surpasses all MASt3R-initialized baselines by a clear margin. Moreover, the two RGBD-based methods rely heavily on high-quality input depth. Since the depth data in ScanNet lacks sufficient precision, they exhibit catastrophic failures across several scenes. Although they perform reasonably well on ScanNet++, they remain inferior to ours even with additional depth information, further confirming the effectiveness of our approach.

Qualitative Comparison. Our method also achieves the best visual performance in qualitative results, as shown in Fig. 4. It produces more accurate geometry, preserves finer scene details, and maintains smoother surfaces on planar regions

Table 2: Ablation study results. From top to bottom: w/o ADC Strategy denotes removing both the iterative adaptive strategy and the supervision of L_α; w/o depth and w/o normal indicate the removal of monocular depth and normal supervision, respectively; w/o densification refers to removing our dense initialization strategy; and w/o MASt3R-SfM replaces our initialization with COLMAP. The first, and second best results in each column are highlighted.

	ScanNet					ScanNet++				
	Acc↓	Comp↓	Prec↑	Recall↑	F1↑	Acc↓	Comp↓	Prec↑	Recall↑	F1↑
w/o ADC Strategy	0.0388	0.0470	0.7900	0.7172	0.7517	0.0333	0.0235	0.8569	0.9003	0.8779
w/o depth	0.0426	0.0469	0.7691	0.7303	0.7491	0.0533	0.0223	0.8493	0.9197	0.8829
w/o normal	0.0398	0.0503	0.7671	0.6878	0.7251	0.0476	0.0282	0.8104	0.8695	0.8385
w/o densification	0.0398	0.0469	0.7740	0.7085	0.7369	0.0357	0.0230	0.8568	0.9117	0.8831
w/o MASt3R-SfM	0.0652	0.0772	0.6917	0.6292	0.6572	0.0438	0.0305	0.7903	0.8507	0.8192
Full Model	0.0374	0.0420	0.7844	0.7352	0.7587	0.0319	0.0227	0.8735	0.9216	0.8967

than other methods. Detailed evaluations of rendering quality and additional qualitative comparisons are provided in the supplementary material.

Computational Cost. For a typical scene with 200 images, MASt3R-SfM incurs approximately 45 minutes of overhead, while depth and normal prior generation take 10 and 15 minutes, respectively. The overall training process requires approximately 40 minutes and consumes around 11 GB of GPU memory. Our DI and ADC strategies do not significantly increase the final Gaussian count (less than 20% increase v.s. 2DGS).

5.3 Ablation Studies

To evaluate the individual contributions of each component in our model, we conducted a series of ablation studies on the ScanNet and ScanNet++ datasets. As shown in Tab. 2, we systematically analyze the impact of the densification strategy, monocular depth, and normal regularization, the dense initialization strategy, and the MASt3R-SfM.

The experimental results clearly demonstrate that high-quality initialization is crucial for effective reconstruction. As seen in the fifth row of Tab. 2, the initialization provided by MASt3R-SfM significantly improves reconstruction performance on indoor scene datasets and constitutes the most substantial contribution. Building upon this, our dense initialization strategy further refines the geometric quality of the reconstruction. Moreover, geometric priors also positively influence surface reconstruction quality. The results in the second and third rows of Tab. 2 show that both monocular depth and normal regularization lead to performance gains, with normal regularization having a particularly notable impact. Although our ADC densification strategy does not yield significant gains in reconstruction metrics, it aids in recovering scene details (Fig. 5). Its effect is mainly visual, as small-scale details typically have minimal influence on the metrics.

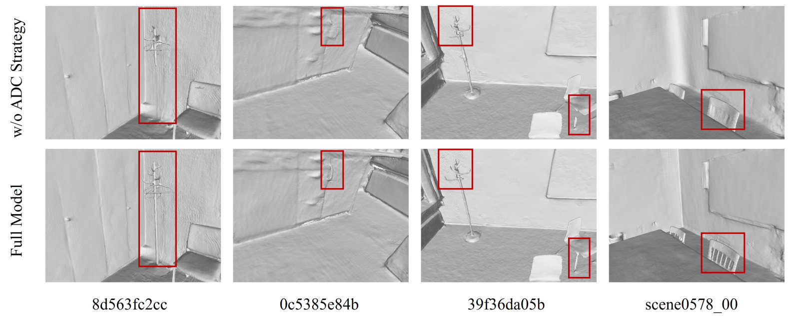

Fig. 5: Comparison of w/o ADC strategy and full_model on ScanNet and ScanNet++ datasets, our method yields finer and more accurate geometric details.

6 Discussion

Although some works such as LM-Gaussian [32] and Matcha Gaussians [11] also leverage monocular depth with advanced SfM models to initialize Gaussian primitives, they are fundamentally different from ours. Effective initialization requires both **accuracy** and **density**. LM-Gaussian targets accuracy by using monocular depth to guide foreground/background segmentation of SfM point clouds and by matching foreground/background depth to resolve multi-coordinate inconsistencies in DUSt3R [26], but **it does not address density**. Matcha Gaussians emphasizes density by converting monocular depth pixels to 3D points and then correcting them with sparse MASt3R-SfM anchors. However, where no SfM point exists, the initialization remains unreliable, yielding **dense but not sufficiently accurate priors**. Our approach reverses this paradigm: starting from the accurate MASt3R-SfM prior, we use monocular depth to guide precise completion, achieving **accurate priors with moderate densification** rather than dense but unstable ones. This distinction is crucial for indoor scenes, where many regions receive limited supervision due to obstructions or camera angles, and inaccurate initialization prevents these Gaussians from converging to correct surfaces. Therefore, our approach is qualitatively and functionally distinct from the referenced methods.

7 Conclusion & Limitation

We propose IndoorSplat, a framework for high-fidelity indoor surface reconstruction. It achieves accurate, complete results using dense initialization from monocular depth, adaptive densification with L_1 opacity regularization, and geometric prior supervision. The limitation of our method is that the densification quality degrades when facing extremely poor monocular depth quality, often caused by motion blur. Additionally, indoor scenes contain rich semantic information, which could be leveraged to further enhance reconstruction quality.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62572436, 62172392), the Major Industrial Tackling Project of Shandong Province (Film and Television Production Industrialization Technology Platform), the Fundamental Research Funds for the Central Universities (Grant No. JB2025028), and the Suzhou Science and Technology Program–Frontier Technology Research Project (Grant No. SYG202327).

References

1. Bao, Y., Ding, T., Huo, J., Liu, Y., Li, Y., Li, W., Gao, Y., Luo, J.: 3d gaussian splatting: Survey, technologies, challenges, and opportunities. IEEE Transactions on Circuits and Systems for Video Technology **35**(7), 6832–6852 (2025)
2. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: Noise reduction in speech processing, pp. 1–4. Springer (2009)
3. Bochkovskiy, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. In: International Conference on Learning Representations. vol. 2025, pp. 75602–75637 (2025)
4. Chen, D., Li, H., Ye, W., Wang, Y., Xie, W., Zhai, S., Wang, N., Liu, H., Bao, H., Zhang, G.: Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. IEEE Transactions on Visualization and Computer Graphics **31**(9), 6100–6111 (2024)
5. Cheng, K., Long, X., Yang, K., Yao, Y., Yin, W., Ma, Y., Wang, W., Chen, X.: Gaussianpro: 3d gaussian splatting with progressive propagation. In: Forty-first International Conference on Machine Learning (2024)
6. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
7. Dai, P., Xu, J., Xie, W., Liu, X., Wang, H., Xu, W.: High-quality surface reconstruction using gaussian surfels. In: ACM SIGGRAPH 2024 conference papers. pp. 1–11 (2024)
8. Dalal, A., Hagen, D., Robbersmyr, K.G., Knausgård, K.M.: Gaussian splatting: 3d reconstruction and novel view synthesis: A review. IEEE Access **12**, 96797–96820 (2024)
9. Duisterhof, B.P., Zust, L., Weinzaepfel, P., Leroy, V., Cabon, Y., Revaud, J.: MASt3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In: International Conference on 3D Vision 2025 (2025), `https://openreview.net/forum?id=5uw1GRBFoT`
10. Fei, B., Xu, J., Zhang, R., Zhou, Q., Yang, W., He, Y.: 3d gaussian splatting as new era: A survey. IEEE Transactions on Visualization and Computer Graphics (2024)
11. Guédon, A., Ichikawa, T., Yamashita, K., Nishino, K.: Matcha gaussians: Atlas of charts for high-quality geometry and photorealism from sparse views. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6001–6011 (2025)

12. Guédon, A., Lepetit, V.: 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. pp. 5354–5363 (2024)
13. Hoppe, H., DeRose, T., Duchamp, T., McDonald, J., Stuetzle, W.: Surface reconstruction from unorganized points. In: Proceedings of the 19th annual conference on computer graphics and interactive techniques. pp. 71–78 (1992)
14. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 conference papers. pp. 1–11 (2024)
15. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the fourth Eurographics symposium on Geometry processing. vol. 7 (2006)
16. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. ACM Transactions on Graphics (ToG) **32**(3), 1–13 (2013)
17. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
18. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. ACM SIGGRAPH Computer Graphics **21**(4), 163–169 (1987)
19. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking. In: 2011 10th IEEE international symposium on mixed and augmented reality. pp. 127–136. Ieee (2011)
20. Ren, X., Turkulainen, M., Wang, J., Seiskari, O., Melekhov, I., Kannala, J., Rahtu, E.: Ags-mesh: Adaptive gaussian splatting and meshing with geometric priors for indoor room reconstruction using smartphones. In: 2025 International Conference on 3D Vision (3DV). pp. 1080–1090. IEEE (2025)
21. Ruan, C., Wang, Y., Guan, T., Zhang, B., Ju, L.: Indoorgs: Geometric cues guided gaussian splatting for indoor scene reconstruction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 844–853 (2025)
22. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
23. Turkulainen, M., Ren, X., Melekhov, I., Seiskari, O., Rahtu, E., Kannala, J.: Dn-splatter: Depth and normal priors for gaussian splatting and meshing. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2421–2431. IEEE (2025)
24. Wang, J., Wang, P., Long, X., Theobalt, C., Komura, T., Liu, L., Wang, W.: Neuris: Neural reconstruction of indoor scenes using normal priors. In: European conference on computer vision. pp. 139–155. Springer (2022)
25. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. Advances in Neural Information Processing Systems **34**, 27171–27183 (2021)
26. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20697–20709 (2024)
27. Xiang, H., Li, X., Cheng, K., Lai, X., Zhang, W., Liao, Z., Zeng, L., Liu, X.: Gaussianroom: Improving 3d gaussian splatting with sdf guidance and monocular cues for indoor scene reconstruction. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 2686–2693. IEEE (2025)

28. Yang, H., Zhang, C., Wang, W., Volino, M., Hilton, A., Zhang, L., Zhu, X.: Improving gaussian splatting with localized points management. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21696–21705 (2025)
29. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. Advances in neural information processing systems **34**, 4805–4815 (2021)
30. Ye, C., Qiu, L., Gu, X., Zuo, Q., Wu, Y., Dong, Z., Bo, L., Xiu, Y., Han, X.: Stablenormal: Reducing diffusion variance for stable and sharp normal. ACM Transactions on Graphics (TOG) **43**(6), 1–18 (2024)
31. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)
32. Yu, H., Long, X., Tan, P.: Lm-gaussian: Boost sparse-view 3d gaussian splatting with large model priors (2024), `https://arxiv.org/abs/2409.03456`
33. Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A.: Mip-splatting: Alias-free 3d gaussian splatting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19447–19456 (2024)
34. Yu, Z., Peng, S., Niemeyer, M., Sattler, T., Geiger, A.: Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. Advances in neural information processing systems **35**, 25018–25032 (2022)
35. Yu, Z., Sattler, T., Geiger, A.: Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. ACM Transactions on Graphics (ToG) **43**(6), 1–13 (2024)
36. Zhang, F., Cao, H., Huang, R.: Consistent quantity-quality control across scenes for deployment-aware gaussian splatting. arXiv e-prints pp. arXiv–2505 (2025)
37. Zhang, W., Xiang, H., Liao, Z., Lai, X., Li, X., Zeng, L.: 2dgs-room: Seed-guided 2d gaussian splatting with geometric constrains for high-fidelity indoor scene reconstruction. arXiv preprint arXiv:2412.03428 (2024)
38. Zhang, Z., Hu, W., Lao, Y., He, T., Zhao, H.: Pixel-gs: Density control with pixel-aware gradient for 3d gaussian splatting. In: European Conference on Computer Vision. pp. 326–342. Springer (2024)