TopoGS: Planar Reconstruction via Topology-aware 3D Gaussian Splatting

Shanshan Pan[1], Jiale Chen[1], Yilin Liu[2], and Hui Huang[1⋆]

[1] Guangdong Provincial Key Laboratory of Visual Media and Multidimensional Intelligence, CSSE, Shenzhen University, China
[2] Autodesk Research, UK

Abstract. Extracting structured, parametric 3D representations from raw images remains a fundamental challenge in computer vision and graphics. While recent advancements in the 3D Gaussian Splatting (3DGS) pipeline integrate planar primitives to yield compact and editable geometry, these approaches typically treat planes as isolated, discrete sets. This lack of topological connectivity hinders robust geometric reasoning, leading to fragmented reconstructions and misaligned boundaries that fall short of the precision for rigorous spatial analysis and professional design workflows. To address this, we introduce TopoGS, the first 3DGS framework to explicitly integrate both planar and topological constraints for coherent 3D reconstruction. Specifically, we extract global 2D topological relationships from multi-view image segmentations and anchor Gaussian primitives to these structural elements. This formulation enables the joint optimization of plane parameters, rendering fidelity, and topological adjacency. By enforcing strict multi-view consistency alongside these topological constraints, our method significantly mitigates geometric misalignments and produces connected, structured 3D models. Extensive evaluations on the ScanNet++ dataset demonstrate that TopoGS achieves state-of-the-art performance, providing a highly robust solution for generating accurate, topologically sound, and visually faithful scene representations.

Keywords: Scene reconstruction · planar reconstruction · 3D Gaussian Splatting · topological constraints

1 Introduction

The creation of high-fidelity, editable 3D representations from raw multi-view images is a cornerstone problem in computer vision and computer graphics. For man-made environments and architectural spaces, the ideal abstraction is a structured planar model, a mathematically compact and interconnected representation explicitly defined by a set of parametric planes, alongside their intersecting lines and corner vertices. However, extracting this pristine structural information directly from unstructured image collections remains exceptionally

⋆ Corresponding author

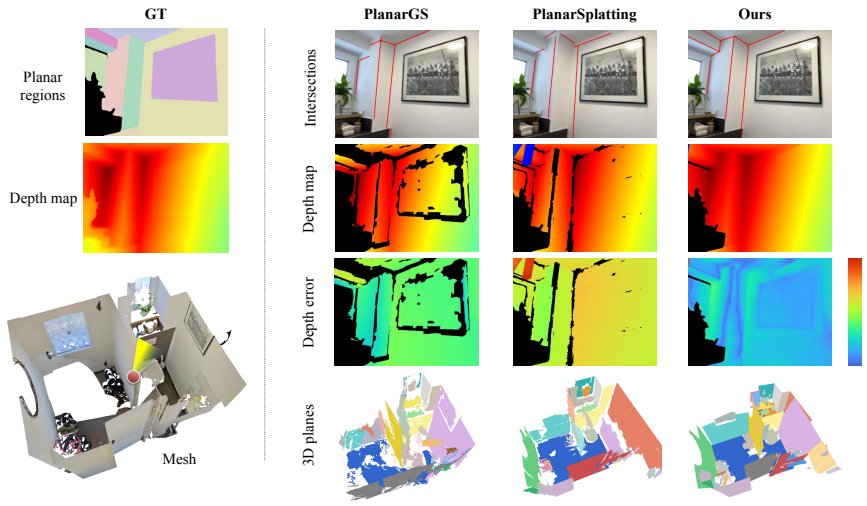

Fig. 1: Motivation of TopoGS. Existing methods show noticeable artifacts in plane intersection projections and depth error maps. In contrast, our TopoGS recovers precise plane boundaries and lower depth errors. This demonstrates that introducing topology constraints acts as a crucial regularizer, providing vital guidance for refining geometric accuracy and effectively suppressing noise in complex scenes. (In the error maps, blue indicates lower error while red signifies higher error.)

challenging. Traditional reconstruction pipelines typically recover scene geometry as unorganized point sets or isolated surface fragments, treating scene elements as discrete entities. Consequently, the critical topological connectivity, denoting the explicit awareness of how these planes intersect, abut, or share boundaries, is frequently lost in the reconstruction process. This fundamental lack of structural awareness hinders robust geometric reasoning, inevitably producing fragmented models and misaligned boundaries that fall short of the precision required for downstream applications such as animation production, robotics, physical simulation, 3D printing, and AR/VR.

Historically, efforts to bridge this structural gap and recover coherent planar models have predominantly relied on a decoupled, two-stage Multi-View Stereo (MVS) pipeline [14] [29] [16] [1] [7]. In the first stage, these methods process multi-view source images to obtain a dense intermediate point cloud. In the second stage, they detect primitive planes from this point cloud and employ complex spatial partitioning to construct a cohesive structural representation. However, the efficacy of this algorithm is fundamentally bottlenecked by the quality of the intermediate geometry. This vulnerability is especially pronounced in indoor scenes, where textureless walls, reflective surfaces, and varying illumination frequently cause both traditional and learning-based MVS techniques to produce noisy, outlier-ridden, or incomplete point clouds. Consequently, the subsequent plane detection becomes highly inaccurate, leading to the omission of critical structural details. Because the final structuring phase relies entirely on this

flawed intermediate representation, it fails to explicitly leverage the rich photometric cues and multi-view consistency present in the raw images, inevitably accumulating severe errors.

To bypass the limitations of intermediate point clouds, recent frameworks increasingly optimize scene representations directly from multi-view images [26] [30] [32] [3] [9] [15]. Notably, 3D Gaussian Splatting (3DGS) [10] has emerged as a highly efficient rendering paradigm. By integrating 3DGS with foundation vision models, recent approaches [21] [6] [32] leverage data-driven monocular priors (e.g., depth, normals, and semantics) to directly recover planar primitives during the optimization. While these 2D cues significantly improve geometric perception in challenging regions, current 3DGS-based methods still fail to achieve the requisite structural awareness. By continuing to treat recovered planes as isolated, independent entities, they remain completely devoid of topological connectivity. However, physical environments are governed by strict structural dependencies, such as the intersecting corners of adjacent walls or the contact boundaries between furniture and floors. We contend that explicitly integrating these topological relationships as vital geometric constraints, rather than relegating them to optional post-processing, can effectively guide the 3DGS optimization landscape, enforce global structural coherence, and dramatically enhance robustness against noise (see Fig. 1).

To this end, we introduce TopoGS, a novel 3DGS framework designed to reconstruct explicitly connected, structured planar models. Our pipeline begins at the image level, extracting 2D region segmentations and constructing their corresponding adjacency graphs directly from the multi-view source images. However, naively lifting these 2D relationships into 3D is fraught with error. Because of perspective occlusions and depth discontinuities, regions that appear adjacent or overlapping in a 2D projection are often mistakenly identified as physically intersecting in 3D. To systematically distill these ambiguous 2D observations into verified 3D topology, we anchor our Gaussian primitives to the segmented regions. By embedding these anchored primitives within the 3DGS optimization loop, we enforce a tri-consistency constraint, integrating photometric, geometric, and topological facets, to jointly refine appearance, scene geometry and plane parameters. This optimization acts as a natural filter, disambiguating genuine physical connections from spurious, view-dependent adjacencies. Finally, we actively enforce these verified 3D adjacencies to assemble the optimized primitives into a globally consistent, structured 3D representation.

Our main contributions are summarized as follows:

- We introduce a novel 3DGS-based framework that integrates geometry and topology constraints to recover consistent plane geometry and topology.
- We propose a method to extract connected structured planar models by leveraging optimized, reliable 3D adjacent relationships.
- Our approach achieves state-of-the-art performance on high-fidelity planar reconstruction tasks within the ScanNet++ dataset.

2 Related Work

2.1 Two-Stage Plane Reconstruction

Traditional approaches for recovering structured planar models typically rely on a decoupled, two-stage paradigm: i) front-end geometry reconstruction to obtain a dense representation (e.g., a point cloud or mesh), followed by ii) back-end plane detection and spatial partitioning to assemble the final structured model.

The primary target of the first stage is to extract raw, dense scene geometry from multi-view source images. Early methods relied on Multi-View Stereo (MVS) [2] [14] [29] [16] [18], which achieves high precision in textured areas but often fails in featureless indoor environments. To mitigate this, neural representations like NeRF and 3DGS incorporate geometric constraints, such as normal regularization and planar priors to enhance surface quality. For instance, PGSR [3] introduces multi-view consistency, while IndoorGS [15] incorporates line and plane priors to bolster indoor reconstruction capabilities. PlanarGS [9] further leverages language-prompt planar priors to provide more reliable geometric guidance. However, these per-scene optimization methods entail heavy computational overhead. Consequently, recent feed-forward approaches [27] [28] [12] [23] [24] leverage Transformer-based attention for cross-view alignment, providing robust geometric priors without iterative optimization.

In the second stage, the target shifts to extracting and structuring semantic primitives from this intermediate geometry. This process typically begins with plane detection via point clustering (e.g., RANSAC [17] or region growing [22]) followed by spatial partitioning (e.g., kinetic or binary partitioning [1] [4] [7]). Despite improvements in front-end geometry, this two-stage workflow suffers from unidirectional error propagation: front-end noise directly misleads back-end plane fitting. Moreover, the plane assembly process cannot back-propagate to refine the front-end geometry, and the rich information in raw images is largely ignored during the second stage, limiting the fidelity of the final planar model.

2.2 Plane Reconstruction from Images

Distinct from two-stage methods, end-to-end planar reconstruction aims to generate plane geometry directly from multiple views. Early research [26] [19] focused on using deep networks to predict planar parameters in voxel space or at the pixel level. With the rise of neural rendering, focus has shifted toward integrating planar primitives into the rendering pipelines of NeRF or 3DGS. For example, NeuralPlane [30] jointly optimizes plane geometry and NeRF representations. PlanarSplatting [21] mimics the alpha-blending capability of 3DGS by proposing plane primitives to replace traditional Gaussian primitives. Similarly, PGS [32] leverages 3D Gaussian primitives for scene modeling but introduces a hierarchical, tree-structured Gaussian Mixture Model (GMM) to group adjacent Gaussian primitives into cohesive planar instances, thereby mitigating the issue of plane over-segmentation. Moreover, GSPlane [6] and GSFlats [20] represent scenes as a hybrid of planes and triangles, categorizing Gaussian primitives

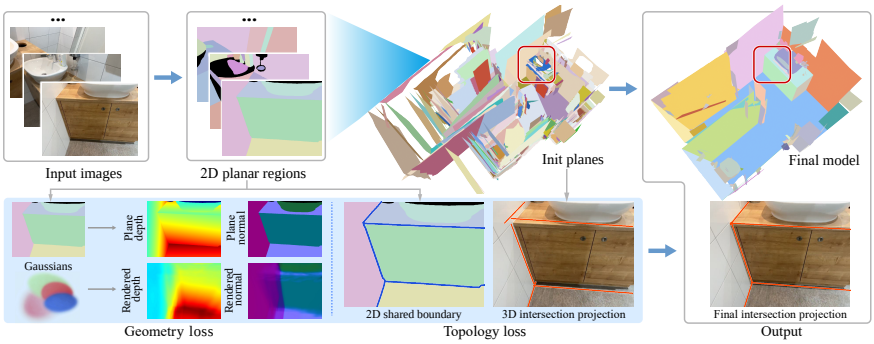

Fig. 2: Overview of TopoGS. (1) **Initialization**: 2D plane regions are extracted from input images and back-projected into 3D space. (2) **Joint optimization**: Plane parameters are refined within the 3DGS-based framework, supervised by our proposed geometry and topology losses. (3) **Structural assembly**: The optimized primitives are integrated into a structured planar model, characterized by topologically consistent and accurate intersection boundaries between adjacent planes.

into planar and non-planar types to adaptively model scenes with both flat and curved surfaces.

While these end-to-end methods successfully enhance local surface smoothness through geometric regularization, they fundamentally treat each plane as an isolated geometric entity. Lacking a global topological graph to explicitly constrain spatial intersections, these approaches often produce misaligned boundaries and fragmented planar shapes, especially under severe occlusion or high noise. Consequently, collaboratively optimizing low-level geometric priors with high-level topological constraints within a unified neural rendering framework remains a critical, unresolved challenge.

3 Method

Given a set of images with camera poses, our goal is to recover the underlying scene structure represented by a connected plane model. We propose TopoGS, a 3DGS-based framework designed to reconstruct a structured model consisting of a set of vertices $\mathcal{V} \subset \mathbb{R}^3$, a set of intersection edges $\mathcal{E} = \{(v_i, v_j) \mid v_i, v_j \in \mathcal{V}\}$, and a set of bounded planes $\mathcal{P} = \{(\pi_k, E_k) \mid E_k \subset \mathcal{E}\}$, where $\pi_k = (\mathbf{n}_k, d_k) \in \mathbb{R}^4$ represents the plane parameters and E_k denotes the index of the bounding edges. As illustrated in Fig. 2, our method comprises three primary stages: (1) Initialization, which extracts plane primitives and their corresponding image-space adjacencies; (2) Joint optimization, which simultaneously refines appearance, scene geometry and plane parameters; and (3) Structural assembly, which converts the optimized primitives into a globally consistent, structured planar model by filtering out false-positive adjacencies.

3.1 Geometry & Topology Initialization

Given a set of images $\{\mathbf{I}_k\}$ and camera poses $\{\mathbf{T}_k\}$, this section aims to recover: i) for each image $\mathbf{I}_k$, a set of 2D planar regions $\{\mathcal{S}_k\}$ and their corresponding plane parameters $\{\pi_i\}$; and ii) a 2D adjacency matrix $\mathcal{A}_{2D}$ that encodes the topological relationships between these regions. These initializations serve as the geometric and topological constraints for subsequent optimization. Following the practices of PlanarSplatting [21], we leverage a pre-trained Metric3D model [8] to predict single-view depth maps $\mathbf{D}_{pred}$ and normal maps $\mathbf{N}_{pred}$ to assist in the initialization.

Hierarchical Segmentation. To ensure that segmentation boundaries align with both semantic logic and geometric consistency, we adopt a two-stage strategy. First, we use the Segment Anything Model (SAM) [11] to obtain initial semantic masks $\mathcal{M}_{sem}$ for each image. These masks provide high-fidelity structural boundaries, effectively delineating critical geometric features such as wall-to-wall junctions. Since a single semantic entity may consist of multiple geometric planes, we further perform Mean-shift clustering [5] within each masked region based on the predicted normal map $\mathbf{N}_{pred}$. This subdivides the masks into a set of planar regions $\{\mathcal{S}_k\}$ for each image, where we preserves the precise edges of SAM while capturing geometric transitions.

World-Space Plane Fitting. To lift 2D planar regions into 3D space, we fit plane parameters π_i for each planar region $s_i \in \mathcal{S}$ using the predicted depth maps $\mathbf{D}_{pred}$. Specifically, for a pixel $p \in s_i$ with homogeneous coordinates $\mathbf{u}_p$, its 3D position $\mathbf{X}_p$ in the world coordinate system is back-projected via:

$$\mathbf{X}_p = \mathbf{R}_k \left(D_{\text{pred}}(p) \cdot \mathbf{K}^{-1} \mathbf{u}_p \right) + \mathbf{t}_k, \tag{1}$$

where $\mathbf{T}_k = [\mathbf{R}_k | \mathbf{t}_k]$ denotes the camera pose (Camera-to-world matrix) and $\mathbf{K}$ is the intrinsic matrix of the camera. We then employ the RANSAC method [17] to estimate the plane parameters from these points.

Adjacency Construction. The initial adjacency graph $\mathcal{A}_{2D}$ inevitably contains spurious connections, such as curved boundaries or segmentation noise. These invalid candidates, which inherently violate the linearity assumption of planar intersections, can be filtered out first to support subsequent optimization. For a pair of adjacent regions (s_i, s_j), we extract their shared boundary pixels $\mathcal{B}_{ij}$ and try to fit a straight line to these pixels. Adjacencies with significant non-linear residuals or insufficient contact length are discarded as they fail to represent valid 3D intersection lines.

The final output includes 2D planar regions $\{\mathcal{S}_k\}$ and a refined 2D adjacency graph $\mathcal{A}_{2D}$. Unlike NeuralPlane [30], our initialization offers: (1) Topological completeness: better 2D adjacency identification via complete segmentation; (2) Boundary accuracy: preservation of sharp SAM edges within the normal-based subdivision; and (3) Robustness: enhanced plane recovery in textureless regions by utilizing predicted depth instead of noisy SFM points.

3.2 Joint Optimization

Directly merging initial plane primitives often fails to yield a coherent scene representation, as it is highly susceptible to multi-view inconsistencies (see Fig. 2, "Init planes"). To address this, we propose an explicit optimization framework built upon 3D Gaussian Splatting (3DGS) to jointly refine both the geometry and topology of the planes.

Plane-Anchored Gaussian Initialization. To facilitate differentiable rendering, we anchor Gaussian primitives to 2D planar regions $\{\mathcal{S}_k\}$. For each view I_k, pixels are sampled on a uniform grid with a stride of $(H/4, W/4)$. For each sampled pixel p, we recover its 3D position $\mathbf{X}_p$ via back-projection using Eq. 1. These 3D coordinates serve as the initial means for the Gaussians, while other attributes are initialized following standard 3DGS protocols.

Loss Functions. We optimize the scene via differentiable Gaussian splatting to synthesize the rendered color $\hat{C}$, depth $\hat{D}$, and normal $\hat{\mathbf{N}}$. In addition to the standard photometric loss $\mathcal{L}_{rgb}$ and the scale loss $\mathcal{L}_{scale}$ proposed in PGSR [3], we introduce two additional key constraints to regularize the optimization.

(1) Geometry Loss ($\mathcal{L}_{geo}$): To ensure that the 3DGS-rendered geometry remains consistent with the parametric planes, we minimize the discrepancy between the rendered depth $\hat{D}(p)$ and the computed plane depth $D_\pi(p)$, similarly to that between the rendered normal $\hat{\mathbf{N}}(p)$ and the plane normal $\mathbf{N}_\pi(p)$. We define the plane depth $D_\pi(p)$ as the orthogonal distance (Z-depth) from the camera center to the intersection point of the ray and the plane. The geometry loss is then formulated as:

$$\mathcal{L}_{geo} = \sum_{p \in \mathcal{I}_k} \lambda_1 \left(|\hat{D}(p) - D_\pi(p)|_1 \right) + \lambda_2 \left(|\hat{\mathbf{N}}(p) - \mathbf{N}_\pi(p)|_1 \right), \tag{2}$$

where λ_1 and λ_2 are weighting factors.

(2) Topology Loss ($\mathcal{L}_{topo}$): To enforce physical closure and structural consistency between two adjacent planes (s_i, s_j), we introduce a topology loss that constrains the projected 3D intersection lines to align with the observed shared boundaries $\mathcal{B}_{ij}$ in the 2D pixel space. For any pair of adjacent regions $(s_i, s_j) \in \mathcal{A}_{2D}$ with corresponding planes $\pi_i = (\mathbf{n}_i, d_i)$ and $\pi_j = (\mathbf{n}_j, d_j)$, their 3D intersection line $\mathbf{L}_{ij}$ is analytically determined by the system of linear equations:

$$\mathbf{n}_i \cdot \mathbf{X} + d_i = 0, \mathbf{n}_j \cdot \mathbf{X} + d_j = 0. \tag{3}$$

We represent $\mathbf{L}_{ij}$ by two 3D points $\{X_0, X_1\}$ and project them onto the image using the camera intrinsic matrix $\mathbf{K}$ and extrinsic parameters $[\mathbf{R}|\mathbf{t}]$ to obtain the 2D projected line $\mathbf{l}_{ij}$ defined by points $\{x_0, x_1\}$. The topology loss is defined as the weighted sum of the mean distances across all adjacent pairs:

$$\mathcal{L}_{topo} = \sum_{(s_i, s_j) \in \mathcal{A}_{2D}} w_{ij} \cdot \mu_{ij}, \tag{4}$$

where $\mu_{ij} = \frac{1}{|\mathcal{B}_{ij}|} \sum_{p \in \mathcal{B}_{ij}} d(p, \mathbf{l}_{ij})$, representing the mean Euclidean distance from all pixels p within $\mathcal{B}_{ij}$ to the ideal line $\mathbf{l}_{ij}$ defined by the vertex pair (x_0, x_1).

Notably, 2D adjacencies $\mathcal{A}_{2D}$ extracted from segmentations are often "spurious", as they may be caused by perspective occlusions or depth discontinuities rather than true physical contact in 3D. To prevent these misleading 2D observations from degrading the 3D geometry, we introduce an uncertainty-aware weighting mechanism w_{ij}:

$$w_{ij} = \exp\left(-\frac{\mu_{ij} + \sigma_{ij}^2}{\tau}\right), \tag{5}$$

where σ_{ij}^2 is the variance of the distances $d(p, \mathbf{l}_{ij})$ for pixels in $\mathcal{B}_{ij}$, and τ is a temperature hyperparameter (set to 50.0). The intuition behind this design is two-fold: i) filtering spurious adjacencies: if two regions are adjacent in 2D but far apart in 3D (high μ_{ij}), or if their shared boundary is geometrically inconsistent with a single straight line (high σ_{ij}^2), the weight w_{ij} decays exponentially and ii) self-driven disambiguation: this allows the optimization loop to act as a natural filter. It prioritizes the refinement of "genuine" physical connections where the 2D pixels already show a strong consensus with the 3D plane intersection, while effectively ignoring view-dependent occlusions that cannot be reconciled with the 3D topology.

Optimization Schedule. We employ a staged strategy to move from coarse geometry to refined topology:

- Stage 1: Coarse optimization (0–4k iterations): Optimize $\mathcal{L}_{rgb}$, $\mathcal{L}_{scale}$, and $\mathcal{L}_{geo}$ to establish a stable geometric foundation.
- Stage 2: Fine optimization (4k–6k iterations): Introduce $\mathcal{L}_{topo}$ to "snap" adjacent planes together and resolve structural ambiguities.
- Adaptive Plane Control: Every 2,000 iterations, we perform two plane operations: merging (combining planes with similar normals/offsets and overlaps) and pruning (removing low-confidence planes) to maintain a compact representation.

The objective function transitions as follows:

$$\mathcal{L} = \begin{cases} \lambda_1 \mathcal{L}_{rgb} + \lambda_2 \mathcal{L}_{scale} + \lambda_3 \mathcal{L}_{geo} & t < 4000, \\ \mathcal{L}_{Stage1} + \lambda_4 \mathcal{L}_{topo} & t \geq 4000. \end{cases} \tag{6}$$

3.3 Structural Assembly

The final stage of our pipeline assembles the optimized parametric planes and their verified topologies into a coherent, structured 3D model. This process transitions from discrete planes to a connected mesh-like representation.

Adjacency Verification. We first refine the initial 2D adjacency graph $\mathcal{A}_{2D}$ to extract a set of verified 3D adjacencies $\mathcal{A}_{verified}$. Benefiting from the joint

optimization in Stage 2, genuine physical connections exhibit high geometric consensus with the 2D shared boundaries $\mathcal{B}_{ij}$. This allows us to apply a strict distance threshold to μ_{ij} to effectively prune spurious adjacencies and retain only true-positive physical intersections.

Intersection Geometry Extraction. For each verified pair in $\mathcal{A}_{verified}$, the 3D intersection line $\mathbf{L}_{ij}$ is analytically derived from the optimized parameters of planes π_i and π_j. To convert these infinite lines into finite edges, we identify shared corners among the 2D boundary segments $\mathcal{B}_{ij}$: if two segments share a common corner, their corresponding 3D lines are deemed adjacent. The 3D corner vertices are then computed as the intersection points of these adjacent lines and used as endpoints to truncate $\mathbf{L}_{ij}$ into bounded segments, forming the skeletal framework of the structured model.

Multi-View Boundary Fusion and Refinement. Since a single perspective often captures only a fragmented portion of a large planar surface, we aggregate the boundary segments of each plane primitive i from all visible multi-view images into a unified global boundary. To ensure structural integrity, we use previously established 3D intersection lines and corner vertices to refine these fused boundaries. This constrained trimming process ensures that adjacent planes converge precisely at their shared geometric anchors, effectively eliminating gaps or overlaps. By combining the refined plane boundaries, intersection lines, and corner vertices, we assemble the primitives into a connected structured model.

4 Experiments

4.1 Dataset, Metrics & Baselines

Datasets. To validate the effectiveness of our proposed framework, we conduct extensive evaluations on ScanNet++ datasets [31], a benchmark widely recognized [30] [21] [9] for its high-fidelity reconstructions and dense annotations. Specifically, we evaluate our method on 25 randomly selected scenes. Following the preprocessing protocols of PlaneRCNN [13], we employ their provided scripts to derive 3D planar ground truth from the raw meshes.

Evaluation Metrics. Following the evaluation framework established by AirPlanes [25], we evaluate our method using four categories of metrics:

- **Geometry:** We report Chamfer Distance (CD) and F-score (5cm threshold) to measure the global geometric alignment with the GT dense mesh.
- **Planar quality:** We evaluate the reconstruction quality of Top-20 largest planes [25] from the ground truth and use the metrics including Planar Fidelity, Accuracy, and Chamfer.
- **Segmentation:** To evaluate 3D plane segmentation, we uniformly sample 200k points on the generated planes to address the disparity in representation density. These points are then used to calculate standard metrics, including Variation of Information (VOI), Rand Index (RI), and Segmentation Covering (SC).

Table 1: Quantitative evaluation on the ScanNet++ dataset. We compare our proposed TopoGS against state-of-the-art two-stage pipelines (combined with AirPlanes' RANSAC extraction) and direct planar reconstruction methods. Top-3 results are highlighted as first, second, and third. Our method achieves leading performance in most planar and geometry metrics, demonstrating the effectiveness of the proposed tri-consistency optimization.

Method	Geometry		Planar			Segmentation		
	CD ↓	F-score ↑	Fidelity ↓	Acc ↓	CD ↓	VOI ↓	RI ↑	SC ↑
PlanarGS [9] + Airplanes	14.02	23.77	24.74	19.25	22.00	3.400	0.926	0.379
DA3 [12] + AirPlanes	8.49	74.40	20.35	12.85	16.60	3.041	0.919	0.431
NeuralPlane [30]	7.56	53.83	14.50	11.83	13.17	2.872	0.944	0.464
PlanarSplatting [21]	9.06	53.83	15.07	14.06	14.56	2.869	0.933	0.463
Ours	6.01	70.18	12.34	9.96	11.15	2.933	0.946	0.463
Ours (+ Assembly)	6.09	69.80	12.96	10.91	11.94	2.867	0.946	0.472

– **Topology:** Since our approach explicitly recovers vertices and edges, we further assess the F-score for these entities, as well as for planes. We apply multiple thresholds (0.05, 0.1, 0.2, 0.5) for a thorough evaluation.

Baselines. We compare our approach against diverse state-of-the-art methods:

– **Two-stage Methods:** These pipelines involve dense reconstruction followed by plane extraction. We select two representative dense reconstruction methods for comparison: (1) PlanarGS [9], a 3DGS-based approach incorporating planar priors, and (2) Depth-Anything-V3 [12], a high-performance feed-forward depth estimation model. For these baselines, we employ the RANSAC-based plane detection protocol provided by AirPlanes [25] to derive planar models from the generated meshes.
– **Direct Planar Reconstruction:** We compare against end-to-end methods that reconstruct planes directly, including NeuralPlane [30] (NeRF-based), and PlanarSplatting [21] (3DGS-based). These methods do not require RANSAC-based post-processing.

For all learning-based baselines, we utilize the officially released pre-trained weights to ensure an equitable comparison.

4.2 Comparisons with baselines

As summarized in Table 1, our proposed TopoGS consistently outperforms both two-stage pipelines and direct planar reconstruction methods across the majority of evaluation metrics. Specifically, in terms of planar reconstruction, TopoGS achieves a significant lead with a Fidelity of 12.34 and an Accuracy (Acc) of 9.96, substantially surpassing the state-of-the-art direct method, NeuralPlane [30]. This performance gain underscores the efficacy of our tri-consistency optimization in refining planar parameters.

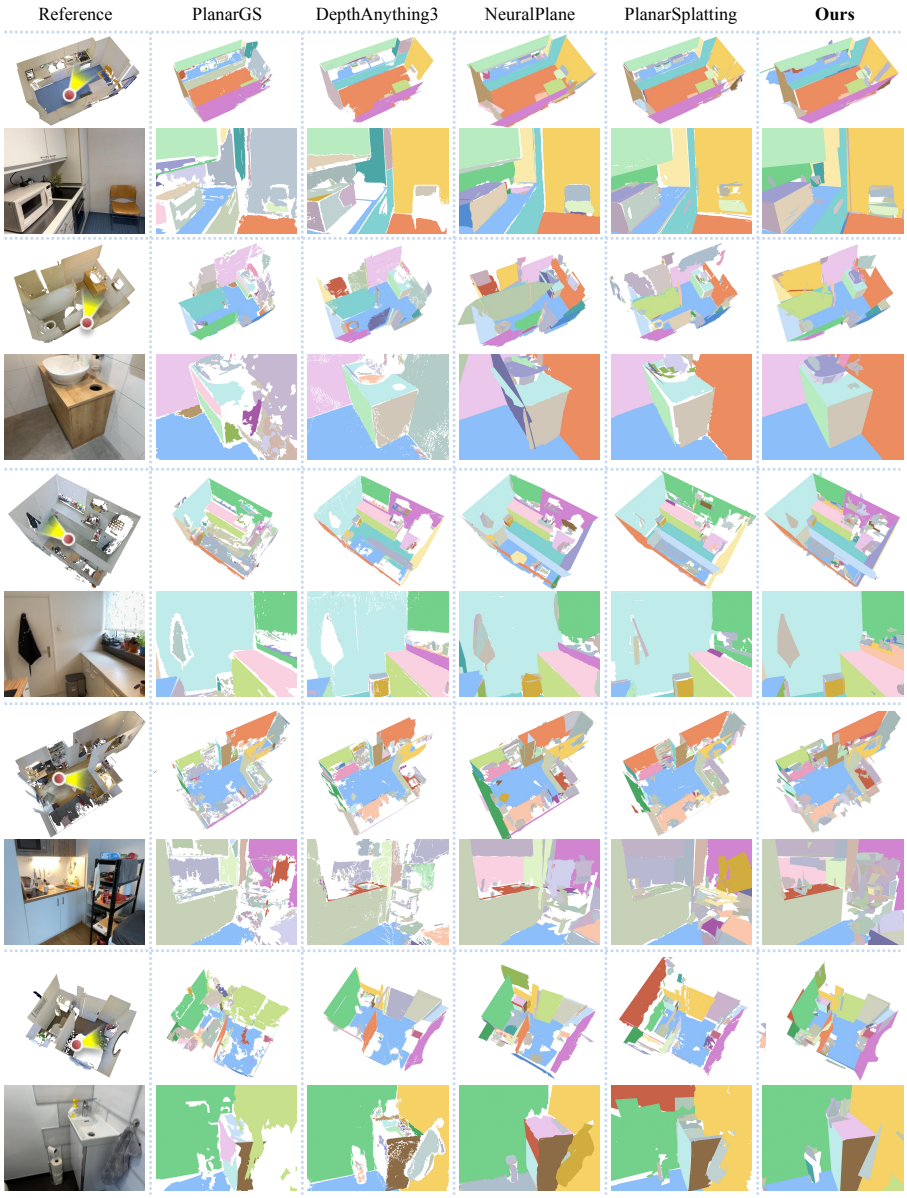

Fig. 3: Qualitative comparison on the ScanNet++ dataset. We evaluate our method against state-of-the-art baselines across diverse environments, including kitchens, bathrooms, and living areas. Compared to existing two-stage and direct reconstruction pipelines, our framework achieves higher structural fidelity and global consistency. Our results exhibit fewer artifacts and more precise boundaries, particularly in challenging regions characterized by weak textures or complex occlusions.

While Depth-Anything-V3 [12] reports a competitive F-score (74.40), attributable to its robust pre-trained depth priors and confidence-based pruning, it fails to recover precise planar structures, as evidenced by its considerably higher planar error metrics. We attribute this deficiency to two factors: first, the difficulty Depth-Anything-V3 faces in textureless or reflective regions, leading to incomplete reconstructions; second, the error propagation inherent in two-stage pipelines where depth estimation and plane extraction are decoupled.

Figure 3 provides a visual comparison between TopoGS and the baseline methods. As illustrated by the qualitative results, TopoGS consistently generates the most structured and visually coherent 3D models across diverse indoor scenes, ranging from cluttered kitchens to complex bathroom fixtures.

Despite the success of planar priors in recovering textureless areas (as demonstrated by NeuralPlane [30], PlanarSplatting [21], and PlanarGS [9]), these methods still struggle with inherent geometric inaccuracies on featureless walls and floors. This occurs because textureless regions introduce photometric ambiguity, settling the optimization into scale-deviant local optima that satisfy planarity and multi-view consistency (e.g., via normalized cross correlation (NCC)) despite being spatially inaccurate. In contrast, our topological constraints capitalize on the semantic understanding of SAM [11]. By utilizing SAM's ability to delineate structural boundaries (such as wall-to-wall intersections) even in textureless regions, our model identifies geometric scaling discrepancies that would otherwise lead to misalignments between projected plane intersections and image segmentation boundaries. Consequently, enforcing these inter-plane topological constraints yields a marked improvement across all metrics.

Furthermore, in challenging scenarios with heavy occlusion, glass reflections, or high sensor noise, Depth-Anything-V3 and PlanarGS often produce inconsistent or noisy depth maps, leading to significant geometric holes. When integrated with AirPlanes [25] for plane extraction, these errors accumulate. Our framework maintains a tighter coupling with the source imagery, combined with topological constraints, effectively resolves ambiguities in these regions and restores a more complete visible range.

Finally, TopoGS excels at recovering sharp, unambiguous boundaries. Unlike baselines that treat the scene as a collection of discrete primitives, TopoGS generates models characterized by physically plausible connections and a "watertight" tendency between adjacent planes. This qualitative superiority reinforces our quantitative findings, demonstrating that explicit topological modeling is indispensable for achieving high-fidelity indoor scene reconstruction.

4.3 Ablation Studies

We conducted an ablation study to evaluate the individual contributions of the geometry loss and topology loss within our framework. The quantitative results are summarized in Table 2.

The exclusion of topological constraints ("w/o Topo") leads to a measurable degradation in performance, with the Chamfer Distance (CD) increasing to 6.54 and the F-score dropping to 64.98. This validates our core insight: explicitly

Table 2: Impact of tri-consistency constraints. This ablation study demonstrates that while photometric consistency provides a baseline, the addition of geometry and topology losses is crucial for improving geometry quality. Our full pipeline significantly outperforms the ablated versions, validating the synergy between the proposed constraints.

Method	Geometry		Planar			Topology (F-score)↑		
	CD ↓	F-score ↑	Fidelity ↓	Acc ↓	CD ↓	V	E	P
w/o Topo	6.54	64.98	13.98	10.96	12.47	19.62	32.05	34.22
w/o Geom	7.56	62.32	16.50	10.66	13.58	12.72	21.28	33.38
Full loss	**6.01**	**70.18**	**12.34**	**9.96**	**11.15**	**21.52**	**35.18**	**36.55**

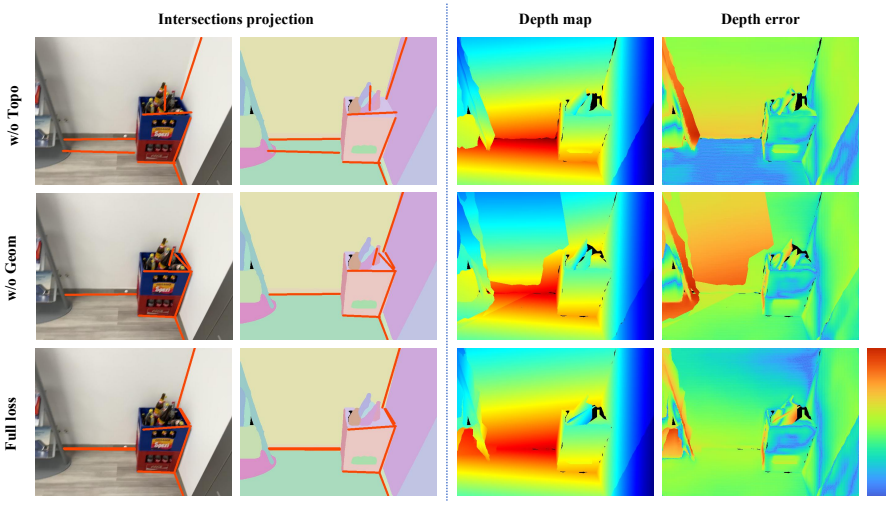

Fig. 4: Qualitative visualization of the ablation study. The "w/o Topo" variant suffers from misaligned planar intersections and high residuals at structural boundaries. While "w/o Geom" improves intersection alignment, it fails to recover accurate absolute depth estimates due to the presence of several floating planes. Our full configuration synchronizes both constraints, achieving sharp, well-aligned boundaries and the lowest overall depth error.

integrating these topological relationships as vital geometric constraints, rather than relegating them to optional post-processing, can effectively guide the 3DGS optimization landscape, enforce global structural coherence (which is evidenced by the improved topology score), and dramatically enhance robustness against noise. The "w/o Geom" variant yields the lowest F-score (62.32) and the poorest Planar Fidelity (16.50), underscoring the necessity of geometric regularization for anchoring planar positions.

Figure 4 visualizes the planar depth errors and the projection errors of plane-to-plane intersections for different variants. A key observation is that the "w/o Geom" configuration achieves deceptively low projection errors but exhibits relatively high 3D depth deviations, indicating that the model overfits to 2D im-

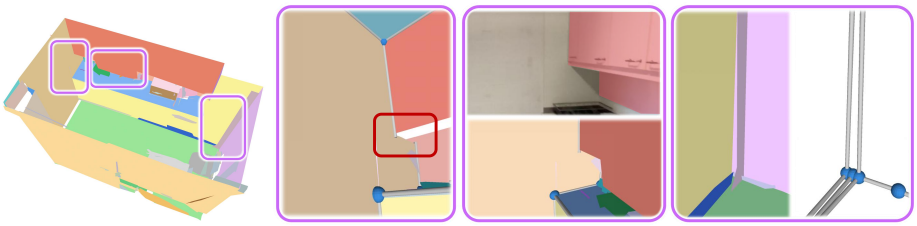

Fig. 5: Limitation. (a) Plane model generated by our method. (b) Incomplete connectivity results in non-watertight gaps at unobserved corners. (c) Inaccurate boundaries resulting from segmentation-induced errors during multi-view fusion. (d) Local inconsistencies generate multiple redundant intersection lines and vertices.

age views without forming a coherent 3D structure. Conversely, while the "w/o Topo" variant maintains lower overall depth error than "w/o Geom," it exhibits significant projection inaccuracies at planar junctions. Notably, even "w/o Topo" depth appears plausible, our full configuration identifies and corrects its subtle geometric discrepancies. These results demonstrate that by jointly optimizing photometric, geometric, and topological consistency, TopoGS effectively synchronizes geometric precision with topological validity, resulting in a consistent high-fidelity scene structure.

4.4 Limitation

It is worth noting that while our optimization encourages tight connectivity, the reconstructed models cannot be strictly watertight. Multi-view images inherently contain occlusions and missing observations, especially at regions which have limited visibility. Consequently, certain intersection lines or vertices may not be fully recovered, leading to non-watertight gaps at unobserved corners (see Fig. 5 (b)). One potential solution to this issue is to leverage geometry and shape priors learned from data to perform reconstruction and generation simultaneously. Another common failure case is due to inaccurate segmentation or complicated local geometry, where we showed in Fig. 5 (c) and (d).

5 Conclusions

In this paper, we present TopoGS, a novel framework for reconstructing structured planar models by integrating geometric and topological constraints within a 3D Gaussian Splatting optimization pipeline. Our method effectively addresses the limitations of existing two-stage and direct planar reconstruction approaches by jointly optimizing plane parameters and their topological relationships. Extensive experiments on the ScanNet++ demonstrate that TopoGS achieves state-of-the-art performance across multiple evaluation metrics, including geometry, planarity, segmentation, and topology. Future work will explore extending our framework to handle more complex scenes with non-planar surfaces and recovering missing regions from incomplete observations.

Acknowledgements

This work was supported in part by Guangdong S&T Program (2024B0101050004), ICFCRT (W2441020), Guangdong Basic and Applied Basic Research Foundation (2023B1515120026), Shenzhen Science and Technology Program (KJZD20240903100022028, KQTD20210811090044003), and Guangdong Provincial Key Laboratory of Visual Media and Multidimensional Intelligence.

References

1. Bauchet, J.P., Lafarge, F.: Kinetic shape reconstruction. ACM Trans. Graph. **39**(5), 1–14 (2020)
2. Bleyer, M., Rhemann, C., Rother, C.: Patchmatch stereo-stereo matching with slanted support windows. In: Brit. Mach. Vis. Conf. vol. 11, pp. 1–11 (2011)
3. 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 Trans. Vis. Comput. Graph. **31**(9), 6100–6111 (2024)
4. Chen, Z., Ledoux, H., Khademi, S., Nan, L.: Reconstructing compact building models from point clouds using deep implicit fields. ISPRS Journal of Photogrammetry and Remote Sensing **194**, 58–73 (2022)
5. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 603–619 (2002)
6. Gan, R., Peng, J., Liu, Y., Luo, C., Li, Q., Zhang, Z.: Gsplane: Concise and accurate planar reconstruction via structured representation. arXiv preprint arXiv:2510.17095 (2025)
7. He, X., Lv, C., Huang, P., Huang, H.: Windpoly: Polygonal mesh reconstruction via winding numbers. In: Eur. Conf. Comput. Vis. pp. 294–311. Springer (2024)
8. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. IEEE Trans. Pattern Anal. Mach. Intell. **46**(12), 10579–10596 (2024)
9. Jin, X., Jin, R., Li, B., Zou, D., Yu, W.: Planargs: High-fidelity indoor 3d gaussian splatting guided by vision-language planar priors. arXiv preprint arXiv:2510.23930 (2025)
10. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Int. Conf. Comput. Vis. pp. 4015–4026 (2023)
12. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
13. Liu, C., Kim, K., Gu, J., Furukawa, Y., Kautz, J.: Planercnn: 3d plane detection and reconstruction from a single image. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4450–4459 (2019)
14. Luo, W., Schwing, A.G., Urtasun, R.: Efficient deep learning for stereo matching. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5695–5703 (2016)

15. Ruan, C., Wang, Y., Guan, T., Zhang, B., Ju, L.: Indoorgs: Geometric cues guided gaussian splatting for indoor scene reconstruction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 844–853 (2025)
16. Sayed, M., Gibson, J., Watson, J., Prisacariu, V., Firman, M., Godard, C.: Simplerecon: 3d reconstruction without 3d convolutions. In: Eur. Conf. Comput. Vis. pp. 1–19. Springer (2022)
17. Schnabel, R., Wahl, R., Klein, R.: Efficient ransac for point-cloud shape detection. In: Comput. Graph. Forum. vol. 26, pp. 214–226. Wiley Online Library (2007)
18. Stier, N., Ranjan, A., Colburn, A., Yan, Y., Yang, L., Ma, F., Angles, B.: Finerecon: Depth-aware feed-forward network for detailed 3d reconstruction. In: Int. Conf. Comput. Vis. pp. 18423–18432 (2023)
19. Sun, J., Xie, Y., Chen, L., Zhou, X., Bao, H.: Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 15598–15607 (2021)
20. Taktasheva, M., Goli, L., Fiorini, A., Li, Z., Rebain, D., Tagliasacchi, A.: 3d gaussian flats: Hybrid 2d/3d photometric scene reconstruction. arXiv preprint arXiv:2509.16423 (2025)
21. Tan, B., Yu, R., Shen, Y., Xue, N.: Planarsplatting: Accurate planar surface reconstruction in 3 minutes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1190–1199 (2025)
22. Verdie, Y., Lafarge, F., Alliez, P.: Lod generation for urban scenes. ACM Trans. Graph. (2015)
23. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5294–5306 (2025)
24. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Scalable permutation-equivariant visual geometry learning. arXiv e-prints pp. arXiv–2507 (2025)
25. Watson, J., Aleotti, F., Sayed, M., Qureshi, Z., Mac Aodha, O., Brostow, G., Firman, M., Vicente, S.: Airplanes: Accurate plane estimation via 3d-consistent embeddings. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5270–5280 (2024)
26. Xie, Y., Gadelha, M., Yang, F., Zhou, X., Jiang, H.: Planarrecon: Real-time 3d plane detection and reconstruction from posed monocular videos. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6219–6228 (2022)
27. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10371–10381 (2024)
28. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
29. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Eur. Conf. Comput. Vis. pp. 767–783 (2018)
30. Ye, H., Liu, Y., Liu, Y., Shen, S.: Neuralplane: Structured 3d reconstruction in planar primitives with neural fields. In: Int. Conf. Learn. Represent. (2025)
31. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Int. Conf. Comput. Vis. pp. 12–22 (2023)
32. Zanjani, F.G., Cai, H., Ackermann, H., Mirvakhabova, L., Porikli, F.: Planar gaussian splatting. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 8905–8914. IEEE (2025)