2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

PIPS: Planar Instance 3D Reconstruction Leveraging Planar Structural Priors

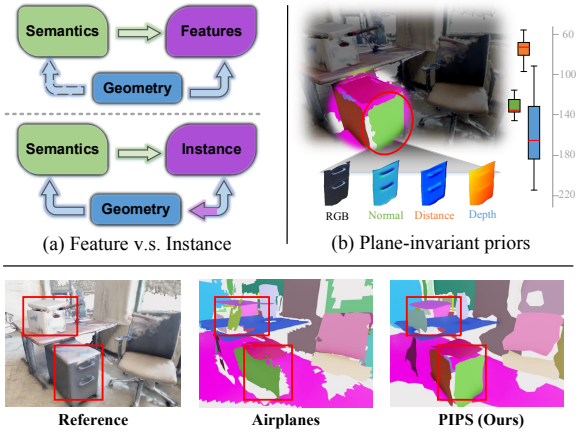
Jiahui Wang, Ye Chen, Yinan Deng, Yi Yang, Yufeng Yue*

***Abstract*— Planar structures, ubiquitous in man-made indoor environments, enable compact and accurate scene abstraction for various downstream tasks. Recent methods distill planar features into learning-based MVS geometries to obtain coherent 3D plane estimation from multi-view inputs. However, the lack of explicit planar instance definitions hinders semantic–geometry alignment, leading to distorted geometry and mismatched semantics. To address this, we propose PIPS, a planar-instance 3D reconstruction method that leverages planar structural priors for both single-view planar segmentation (SGPS module) and multi-view instance association (MVPI module). The planar instance point clouds are regularized by planar distances and then converted into complete planar meshes via an instance-level planar meshing strategy. Extensive experiments on hundreds of indoor scenes demonstrate the superior performance of our method, which is less dependent on annotations and requires no feature optimization. The effectiveness of each component is further verified through comprehensive ablation studies. The project page of PIPS is available at https://pips325.github.io.**

I. Introduction

Planar structures are ubiquitous in man-made indoor environments, ranging from large-scale elements such as walls and floors that form the framework of a room [1], to smaller surfaces such as tabletops and cabinets that carry the functional objects of interest [2]. Such compact and efficient planar abstraction of environments plays a vital role in robotic applications like navigation [2] and virtual reality [3], [4]. These applications require consistent 3D planar reconstruction at scene-level in multi-view configurations [5].

Over the past few years, 2D planar segmentation has been extensively studied [6]–[10] and has recently achieved impressive progress with the adoption of Vision Transformers (ViT) [11], [12]. However, inconsistent planar segments and geometric estimation across views hinder their direct application in multi-view settings that process sequential image inputs. Conventional 3D planar segmentation methods [13]–[15], which fit planes on known 3D geometry (e.g. point cloud) using RANSAC [16], frequently produce semantically fragmented planes and remain highly sensitive to noise [5]. With the rapid development of multi-view stereo (MVS) [17]–[19], recent researchers [5], [20]–[22] have begun to explore how inconsistent 2D planar segments can be distilled into multi-view consistent 3D geometry. This is typically achieved by learning-based MVS framework such as Neural Radiance Field (NeRF [19]) and 3D Gaussian Splatting (3DGS [18]), where conditioning on planar features enables the extraction of coherent 3D planes.



(a) Feature v.s. Instance (b) Plane-invariant priors

(c) Showcasing Results of 3D Plane Reconstruction

Fig. 1. Motivations of PIPS. (a) Feature-based vs. instance-based planar reconstruction pipelines. (b) Normal, distance, and depth value distributions within a plane; narrower gray-value boxplots indicate stronger planar invariance. (c) A detailed comparison of planar reconstruction on the ScanNet dataset, showing that our method yields more complete and aligned planes.

However, existing approaches overlook an inherent property of planar representation: ***planar regions are naturally partitioned by geometric structures, and structural attributes such as normals and distances remain highly consistent within each plane.*** Current multi-view methods [5], [20], [21] typically treat MVS geometry merely as a medium for distilling planar segmentation IDs, without enforcing sufficient geometric constraints on the underlying planes. Consequently, planar semantics and geometry often become misaligned, leading to distorted planes with inaccurate semantics, as illustrated in Fig. 1(c). More fundamentally, their implicit planar features cannot be consistently anchored to specific planes during distillation, which reduces reconstruction flexibility and increases the training burden.

Therefore, to explicitly establish a bridge between planar semantics and geometry, we draw inspiration from the recent lightweight 3D instance segmentation framework [23] and extend the notion of instances from the entity level to the planar level (Fig. 1(a)). To further enforce alignment between semantics and geometry across both 2D and 3D domains, we incorporate planar structural priors [24]–[26], such as surface normals and camera-view distances in Fig. 1(b), for both 2D planar segmentation and 3D instance association. Building on this, we propose **PIPS**, a **P**lanar **I**nstance 3D reconstruction method Leveraging **P**lanar **S**tructural Priors and consists of the following three modules:

The work is supported by National Natural Science Foundation of China under Grant 62473050, 92370203, 625B2024, Beijing Natural Science Foundation Undergraduate Research Program QY25270. (*Corresponding author: Yuefeng Yue, yueyufeng@bit.edu.cn)

[1] School of Automation, Beijing Institute of Technology, Beijing, China.

1) Single-view initialization: The Structure-Guided Planar Segmentor (SGPS) utilizes structural priors from monocular models to partition 2D planar segments and estimate coarse geometry. 2) Multi-view Association: Then the Multi-View Planar Integrator (MVPI) constructs a mask graph to associate cross-view segments with consistent instance IDs, generating planar instance point clouds with smooth 2.5D geometry. 3) Planar Mesh Extraction: An instance-level planar meshing is performed on each distance-normalized planar instance point cloud yielding final 3D planar reconstructions. Our contributions are summarized as follows:

- We propose PIPS, a planar instance 3D reconstruction method that bridges planar semantics and aligned geometry using planar structural priors.
- We introduce a Structure-Guided Planar Segmentor (SGPS) that detects planar regions from monocular cues without annotations and benefits from sparse labels.
- We introduce a Multi-View Planar Integrator (MVPI) that leverages multi-view planar priors for consistent 2D-to-3D planar instance association and regularization.
- Our method achieves superior performance on 3D planar reconstruction in hundreds of public indoor scenes.

II. Related Works

A. Planes from single-view images

Many learning-based methods [6]–[10] treat single-view plane reconstruction as a plane instance segmentation task, typically supervised with manual annotations. Among this, PlaneNet [7] first adopt Deep Neural Network (DNN) to directly infer planar parameters and masks, while PlaneRCNN [6] extends Mask R-CNN [27] to estimate depth, normals, and offsets for planar regions. Recent works leverage query-based Vision Transformers to achieve state-of-the-art single-view results. PlaneRecTR [11] unifies all subtasks of single-view plane recovery via query-based modeling, improving prediction consistency, while ZeroPlane [12] achieves zero-shot generalization through large-scale multi-dataset training. However, the inconsistency in segmentation and metric scale across frames poses a major challenge for applying single-view planar methods in multi-view reconstruction [5].

B. Planes from multi-view images

Building upon single-view methods, sparse-view approaches aim to directly predict plane correspondences between typically two posed views, using either end-to-end learning [28], [29] or a RANSAC-based paradigm [30]. Yet, they still lack global association, limiting the reconstruction of entire 3D scenes from monocular video inputs. To address this, PlanarRecon [31] progressively learns a 3D planar volume from posed RGB inputs, from which oversimplified 3D planes are extracted.

Due to the high cost of 3D plane annotations [31], recent methods utilize multi-view neural 3D reconstruction to lift single-view 2D plane segments into 3D, typically through 2D-to-3D planar feature distillation followed by similarity-based plane grouping. Airplanes [5] trains a per-scene MLP to fuse multi-view 2D planar features into a 3D TSDF estimated by [17], which is subsequently clustered into 3D planes. However, the decoupling of geometry recovery and plane estimation bottlenecks the overall performance [20].

To couple this, NeuralPlane [20] represents scenes as a set of local planar primitives, whose geometry is supervised by corresponding plane observations and further grouped into 3D planes via a push-pull feature loss. Limited by the implicit nature of Neural Radiance Field (NeRF [19]), [20] struggles to constrain explicit geometry accurately, resulting in low-resolution, patch-like planar decompositions. Building on recent advances in 3D Gaussian Splatting (3DGS [18]), PlanarSplatting [32] fits super-resolution 3D planes using customized rectangular Gaussian primitives.

However, without an explicit definition of planar instances, the above methods only achieve local planar regularization with misaligned semantics. In contrast, our PIPS achieves 3D planar-instance segmentation while enforcing planarity regularized by multi-view planar structural priors.

C. 3D Instances from multi-view images

With the rise of powerful 2D instance segmentation models like Segment Anything (SAM) [33], a common strategy for 3D instance segmentation is to associate cross-view masks via depth-based projection into 3D space [34]–[37]. SAM3D [38] projects SAM-predicted masks into 3D point clouds then applies iterative bottom-up merging. Unlike expensive point cloud processing, MaskClustering [23] builds a view consensus based mask graph to cluster entity-level 2D masks [39] into 3D-consistent instances in a lightweight manner. OpenObj [40] further enhances the clustering process by incorporating semantic similarities between masks, derived from vision-language embeddings [41]. However, both the above 2D foundation models and association mechanisms primarily focus on object entities, rather than planar instances, which are typically characterized by stricter geometric constraints and more ambiguous semantics.

III. Method

Given a sequence of RGB image frames $\mathcal{C} = \{\boldsymbol{C}_t\}$ and their corresponding camera poses $\mathcal{P} = \{\boldsymbol{P}_t\}$, PIPS introduces a planar instance 3D reconstruction method that aligns planar semantics and geometry. Building on the structural prior: ***"Pixels from the same plane share the similar Normals and Distances"***, we present the proposed 3D planar reconstruction method PIPS through the following three sections: the SGPS module (Sec. III-A) for single-view planar abstraction, the MVPI module (Sec. III-B) for multi-view planar aggregation. Finally, an instance-level planar meshing strategy is employed for the final reconstruction (Sec. III-C). The detailed pipeline is illustrated in Fig. 2

A. SGPS: Structure-Guided Planar Segmentor

Leveraging the planar structural priors in Fig. 2, we design an annotation-free 2D planar segmentation module based on cues from Metric3D [42] and primitives from SAM [33].

Normal-Guided Planar Clustering. To mitigate the plane-agnostic behavior of SAM, we incorporate a normal-driven

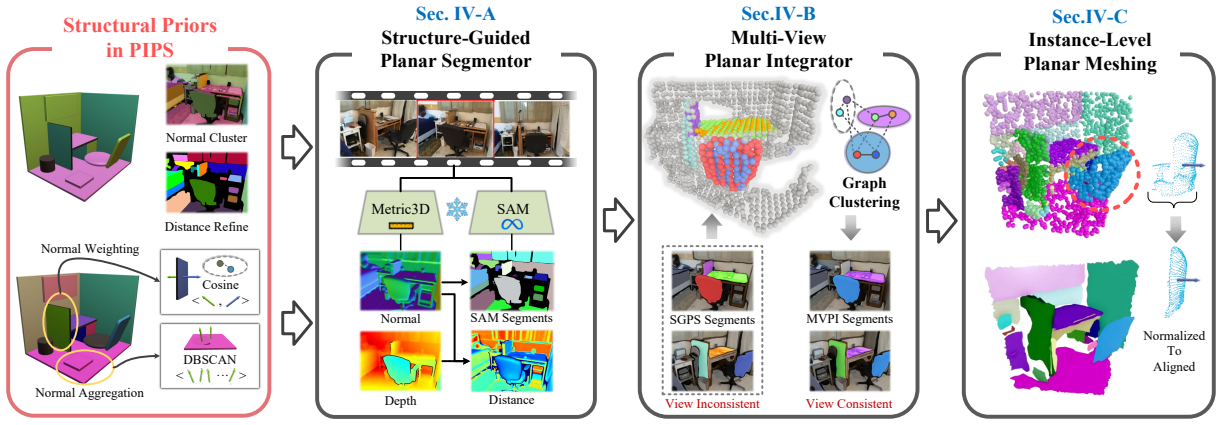


Fig. 2. Planar structural priors are utilized throughout the entire PIPS pipeline. In the SGPS module, normal and distance priors are first employed for single-view planar segmentation. Next, the MVPI module aggregates these segments into planar-instance point clouds via planar-normal guided mask clustering. Finally, an instance-level meshing is applied to each distance-normalized instance for semantically and geometrically aligned planar reconstruction.

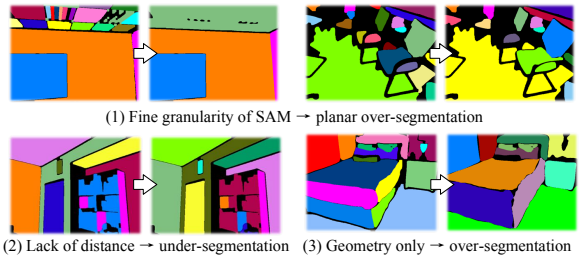


Fig. 3. Challenges faced by the SGPS module (Left: failure, Right: ours).

clustering mechanism to pre-partition potential planar regions before invoking $SAM(\cdot)$. Given monocular predictions from Metric3D, we obtain surface normals and depths as $\hat{\boldsymbol{N}} = \{\hat{\boldsymbol{N}}_t\}$ and $\hat{\mathcal{D}} = \{\hat{\mathcal{D}}_t\}$. For each frame, the predicted normal map $\hat{\boldsymbol{N}}_t$ is grouped using a cuML-accelerated KMeans algorithm, yielding dominant-normal masks $\{M_{t,n}^{norm}\}$. These masks constrain SAM to produce coarse planar candidates $\{M_t^{sam}\}$ with geometry-aware initialization. The overall procedure is summarized as follows:

$$M_{t,n}^{norm} = KMeans(\hat{\boldsymbol{N}}_t, n_{cluster} = 5) \tag{1}$$

$$M_t^{sam} = \sum SAM(\mathbf{C}_t, M_{t,n}^{norm}), n = 1, 2, \ldots, 5 \tag{2}$$

Distance-Guided Planar Refinement. In contrast to the fine-grained SAM masks adopted in [20], which emphasize part-level decomposition, we intentionally employ coarser masks to preserve planar completeness and improve efficiency (see Fig. 3(1)). However, without spatial distance awareness, such coarse segmentation may cause planar under-segmentation, e.g., merged cabinet sides in Fig. 3(2). To alleviate this issue, we incorporate camera-view distance as a refinement cue, estimated from monocular normals and depths

$$\hat{\boldsymbol{D}}(\boldsymbol{p}) = \hat{\boldsymbol{N}}(\boldsymbol{p})\boldsymbol{K}^{-1}\hat{\mathcal{D}}(\boldsymbol{p})\tilde{\boldsymbol{p}} \tag{3}$$

where $\boldsymbol{p} = [u, v]^T$ denotes pixel coordinates in the image plane, $\tilde{\boldsymbol{p}}$ their homogeneous form, and $\mathbf{K}$ the camera intrinsic matrix. Refinement operates on connected components within each planar candidate M_t^{sam}, splitting regions with

Algorithm 1: Distance-Guided Planar Refinement

Input: Normal $\hat{\boldsymbol{N}}_t$, depth $\hat{\mathcal{D}}_t$, distance $\hat{\boldsymbol{D}}_t$, image $\boldsymbol{C}_t$
Output: Refined planar mask M_t^{ref}

Cluster $\boldsymbol{C}_t$ with $\{M_{t,n}^{norm}\}$ and obtain planar candidates M_t^{sam} via Eq. 2;
Initialize $M_t^{ref} \leftarrow 0$, $\gamma \leftarrow 1$;
foreach $m_s^{sam} \in M_t^{sam}$ **do**
  Extract connected components $\{m_{s,k}^{sam}\}$; Initialize $\mathcal{L}_s^{dist}, \mathcal{L}_s^{id} \leftarrow \emptyset$;
  foreach $m_{s,k}^{sam}$ **do**
    Compute mean distance $D_{s,k}^{mean}$;
    if $\mathcal{L}_s^{dist} \neq \emptyset$ **then**
      $\boldsymbol{D}_{s,k}^{diff} = |D_{s,k}^{mean} - \mathcal{L}_s^{dist}|$;
      if $\max(\boldsymbol{D}_{s,k}^{diff}) > 0.2 D_{s,k}^{mean}$ **then**
        $\gamma \leftarrow \max(\mathcal{L}_s^{id}) + 1$;
      else
        $\gamma \leftarrow \mathcal{L}_s^{id}[\arg\min(\boldsymbol{D}_{s,k}^{diff})]$;
      end
    end
    Append $D_{s,k}^{mean}$ to $\mathcal{L}_s^{dist}$ and γ to $\mathcal{L}_s^{id}$;
    Assign $M_t^{ref}[m_{s,k}^{sam}] \leftarrow \gamma$;
  end
end
return M_t^{ref}

inconsistent camera-view distances to produce refined masks M_t^{ref}. Algorithm 1 summarizes the procedure.

Sparse Planar Label Fusion. Although distance refinement produces dense and well-separated planar masks M_t^{ref} without semantic supervision, SAM-based segmentation remains prone to over-segmentation (e.g., splitting bed sides and quilts in Fig. 3(3)). To regularize such fragmentation, we integrate sparse yet structurally reliable planar segments M_t^{xpd} predicted by X-PDNet [8]. The resulting SGPS planar segments $\{M_{t,i}^{sgps}\}$ define a per-pixel label map $M_t^{sgps}(\boldsymbol{p})$:

$$M_t^{sgps}(\boldsymbol{p}) = \begin{cases} M_t^{xpd}(\boldsymbol{p}), & \text{if } M_t^{xpd}(\boldsymbol{p}) \neq 0, \\ M_t^{ref}(\boldsymbol{p}), & \text{if } M_t^{xpd}(\boldsymbol{p}) = 0, \end{cases} \tag{4}$$

SGPS Planar Structural Outputs. SGPS produces single-view planar segments (Fig. 2) and the point cloud $\mathcal{P}_I$ for MVPI. For each segment $M_{t,i}^{sgps}$, we cluster per-pixel normals $\hat{\boldsymbol{n}}_{\boldsymbol{p}}$ via agglomerative clustering [43] and select the dominant cluster with the largest support. The segment-level orientation $\boldsymbol{N}_{t,i}^{sgps}$ is obtained by averaging normals within this cluster:

$$n^* = arg\max_n |\mathcal{C}_{i,n}| \tag{5}$$

$$\boldsymbol{N}_{t,i}^{sgps} = \frac{1}{|\mathcal{C}_{i,n^*}|} \sum_{\boldsymbol{p}\in\mathcal{C}_{i,n^*}} \hat{\boldsymbol{n}}_{\boldsymbol{p}} \tag{6}$$

The normalized planar normal is then paired with its mask to form the SGPS planar segment $\boldsymbol{S}_{t,i}^{sgps} = \{M_{t,i}^{sgps}, \boldsymbol{N}_{t,i}^{sgps}\}$.

To enable multi-view planar association in MVPI, we adopt a meshing–rendering–projection pipeline. We reconstruct a coarse mesh from $(\hat{\mathcal{D}}_t, \boldsymbol{P}_t)$ via TSDF-Fusion [17], render a clean depth map $\mathcal{D}_t^{re}$ using the ScanNet++ renderer [44], and project it across views to form the SGPS point cloud $\mathcal{P}_I$. This process suppresses monocular depth noise and provides a stable geometric basis for cross-view association.

B. MVPI: Multi-View Planar Integrator

To obtain 2D–3D consistent identities and instance-level planar normals for mesh extraction, we integrate the SGPS segments $\boldsymbol{S}_{t,i}^{sgps}$ with the SGPS point cloud $\mathcal{P}_I$. This produces multi-view consistent planar segments $\boldsymbol{S}_{t,j}^{mvpi}$ and the 3D instance point cloud $\mathcal{P}_{II}$. We first construct a view-consensus weighted mask graph and perform clustering [23].

Original View-Consensus based Mask Clustering. Following the principle that two masks belonging to the same instance should be jointly observed by the same mask in a sufficient number of frames, [23] introduces a plug-and-play criterion called View-Consensus $c(.,.)$ between $M_{t',i'}^{sgps}$ and $M_{t'',i''}^{sgps}$:

$$c(M_{t',i'}^{sgps}, M_{t'',i''}^{sgps}) = \frac{|\mathcal{M}(M_{t',i'}^{sgps}) \cup \mathcal{M}(M_{t'',i''}^{sgps})|}{|\mathcal{F}(M_{t',i'}^{sgps}) \cup \mathcal{F}(M_{t'',i''}^{sgps})|} \tag{7}$$

more details about the visible frame set $\mathcal{F}(M_{t',i'}^{sgps})$ and contained mask set $\mathcal{M}(M_{t',i'}^{sgps})$ please refer to [23].

Then the view-consensus rate weighted mask graph $G = (N, E)$ can be constructed, where each node in N denotes a SGPS mask $M_{t,i}^{sgps}$, and each edge is created between two masks whenever their view-consensus exceeds 0.9, representing current graph as an adjacency matrix $A(G) \in \mathbb{R}^{|N|\times|N|}$, where $|N|$ denotes the total number of nodes:

$$A(G_k)_{ij} = \mathbf{1}(\boldsymbol{c}(M_i^{sgps}, M_j^{gphs}) \geq 0.9) \tag{8}$$

Once the iterative clustering converges, each node $\mathcal{C}_j^{mvpi}$ is expected to contain all multi-view SGPS masks belonging to the same instance I_j^{mvpi}, along with their corresponding instance point cloud P_j^{mvpi}.

Planar Normal guided Mask Clustering. Since mask consensus mainly relies on point-cloud distance, planar association becomes ambiguous, as illustrated in Fig. 4, where

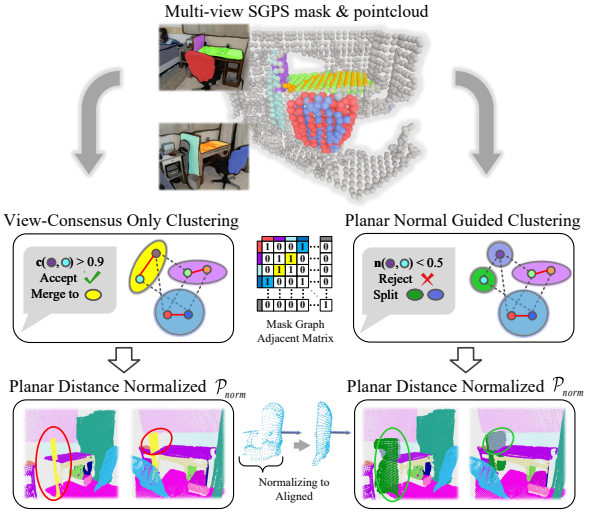


Fig. 4. Comparison between the view-consensus and planar-normal guided mask clustering during MVPI.

the two sides of a desk panel belong to different planes. To resolve this, we introduce SGPS planar-normal similarity as an additional cue, forming the Planar Normal-guided Mask Clustering.

Specifically, each planar normal $\boldsymbol{N}_{t,i}^{sgps}$ is first transformed into the world coordinate system using the frame pose $\boldsymbol{P}_t = [\boldsymbol{R}_t, \boldsymbol{t}_t]$, yielding $\boldsymbol{N}_{t,i}^{sgps_w} = \boldsymbol{N}_{t,i}^{sgps} \cdot \boldsymbol{R}_t{}^T$. Then normal similarity between SGPS masks can be computed as:

$$mn(M_{t',i'}^{sgps}, M_{t'',i''}^{sgps}) = \frac{\boldsymbol{N}_{t',i'}^{sgps_w} \cdot \boldsymbol{N}_{t'',i''}^{sgps_w}}{|\boldsymbol{N}_{t',i'}^{sgps_w}| \cdot |\boldsymbol{N}_{t'',i''}^{sgps_w}|} \tag{9}$$

Accordingly, we construct a mask-normal similarity adjacency matrix $N(G_k) \in \mathbb{R}^{|N|\times|N|}$ based on $A(G_k)$, where

$$N(G_k)_{ij} = \mathbf{1}(\boldsymbol{n}(M_i^{sgps}, M_j^{gphs}) \geq 0.5) \tag{10}$$

The new adjacency matrix is then updated as

$$A(G_k) = A(G_k) \cap N(G_k) \tag{11}$$

During iterative clustering, when nodes are merged, the normal of the new mask $\boldsymbol{N}^{new}$ is updated via dominant clustering over the normal set $\{\boldsymbol{N}_{t_s,i_s}^{sgps_w}\}$, analogous to Eq. 5. The resulting node $\mathcal{C}_j^{mvpi}$ is then decomposed into multi-view consistent MVPI masks $\{M_{t,j}^{mvpi}\}$, where $j \in \mathcal{J}$ denotes the cross-view instance index.

MVPI Planar Structural Outputs. These include the multi-view planar segments in Fig. 2 and the planar-instance point cloud $\mathcal{P}_{II}$. The dominant instance-level normal $\boldsymbol{N}_j^{mvpi}$ from the converged mask graph is transformed to each frame t containing the instance, yielding per-view normals $\{\boldsymbol{N}_{t,j}^{mvpi} = \boldsymbol{N}_j^{mvpi} \cdot \boldsymbol{R}_t\}$. Combined with $\{M_{t,j}^{mvpi}\}$, this produces planar segments $S_{t,j}^{mvpi}$ for MVS supervision. The clustered normal $\boldsymbol{N}_j^{mvpi}$ is further assigned to the corresponding point cloud P_j^{mvpi}, forming the instance-level point cloud $\mathcal{P}_{II}$ for subsequent planar mesh extraction.

C. Instance-Level Planar Meshing

Since ball-pivoting meshing [45] operates directly on point clouds, applying it to each planar set $P_j^{mvpi} \in \mathcal{P}_{II}$ produces the final planar instance meshes $\mathcal{R}_{plane}$. However, as $\mathcal{P}_{II}$ is decomposed from the SGPS point cloud $\mathcal{P}_I$ without strict planar constraints, non-planar geometry may arise (e.g., the chair in Fig. 4). We therefore regularize the planarity of P_j^{mvpi} via internal distance normalization before meshing. Specifically, the normalized set P_j^{norm} is obtained by projecting each point of the RANSAC-filtered P_j^{mvpi} onto its mean-distance plane:

$$D_j^{aver} = Mean(\boldsymbol{D}_j^{mvpi_w} = P_j^{mvpi} \cdot \boldsymbol{N}_j^{mvpi}) \tag{12}$$

$$P_j^{norm} = P_j^{mvpi} - (\boldsymbol{D}_j^{mvpi_w} - D_j^{aver}) \otimes \boldsymbol{N}_j^{mvpi} \tag{13}$$

where D_j^{aver} denotes the dominant plane position of the point cloud, and $\otimes$ applies the offset along the normal to each point. The corresponding planar meshes $\mathcal{R}_j$ are then extracted and collectively form the scene-level planar reconstruction $\mathcal{R}_{plane}$.

IV. Experiments

We comprehensively evaluate PIPS on hundreds of indoor scenes from two high-fidelity public datasets. The evaluation includes quantitative comparisons using planar semantic and geometric metrics widely adopted in previous baselines. Moreover, we conduct extensive ablation studies on each module and its components to validate their effectiveness.

A. Experimental Setup

Implementation Details: Our implementation is based on PyTorch and evaluated on a single RTX 4090 GPU. To balance accuracy and efficiency, we uniformly sample and process no more than 1,000 images per scene.

Dataset: For ScanNetV2 [46], we follow the standard protocol [5] and evaluate on 100 test scenes. For ScanNet++ [44], we randomly select 30 scenes from the official *nvs_sem_val* split. For evaluation, we use the scripts provided by PlanarRCNN [6] to generate ground-truth 3D plane annotations.

Baseline: To comprehensively benchmark PIPS for planar reconstruction, we compare with two categories of methods.

(I) 3D geometry-based methods. These approaches require no semantic annotations and extract planes directly from reconstructed meshes via geometric clustering. We consider: (1) the pretrained MVS method [17] (SR+Ransac); (2) TSDF-fusion meshing from [42] (M3D+Ransac) and (3) the 3DGS-based high-fidelity reconstruction method [25] (PGSR+Ransac). All reconstructions follow their official implementations, while sequential RANSAC is implemented following the standard setup in [5].

(II) Multi-view learning-based methods. These approaches learn planar semantics and geometry from multi-view planar segmentation or 2.5D observations. We compare against all open-source methods, including the pioneering PlanarRecon [31], the semantic method Airplanes [5], the NeRF-based NeuralPlane [20], and the 3DGS-based PlanarSplatting [32].

TABLE I. Quantitative Results on the 100 test splits of ScanNetV2. The top 3 results are highlighted as first, second, and third.

Method	VOI↓	RI↑	SC↑	Cham.p.↓	Cham.a.↓	F-score↑
SR+Ransac	2.583	0.944	0.498	10.444	5.650	63.213
PGSR+Ransac	2.784	0.947	0.461	12.545	6.476	57.251
M3D+Ransac	2.199	0.958	0.562	7.978	4.161	73.843
PlanarRecon	3.252	0.916	0.394	18.267	10.126	42.583
Airplanes	2.248	0.959	0.573	8.394	5.390	63.743
Neuralplane	3.824	0.918	0.301	12.986	6.579	58.334
PlanarSplatting	2.502	0.948	0.532	9.200	4.830	68.850
PIPS (Ours)	2.134	0.959	0.594	7.234	4.322	74.165

TABLE II. Quantitative Results on the 30 random scenes of ScanNet++. The top 3 results are highlighted as first, second, and third.

Method	VOI↓	RI↑	SC↑	Cham.p.↓	Cham.a.↓	F-score↑
SR+Ransac	3.602	0.950	0.355	22.089	11.557	33.670
PGSR+Ransac	3.455	0.952	0.387	18.418	8.668	49.460
M3D+Ransac	2.919	0.948	0.457	15.937	8.467	48.895
PlanarRecon	3.966	0.930	0.322	27.905	15.996	32.495
Airplanes	3.055	0.951	0.440	19.386	11.479	34.895
Neuralplane	4.597	0.936	0.233	19.874	11.138	42.123
PlanarSplatting	3.199	0.942	0.436	16.929	8.809	52.319
PIPS (Ours)	2.820	0.952	0.483	14.798	7.493	53.809

For fair comparison, we disable the post-processing module in [20] to avoid potential ground-truth leakage.

Metrics: For planar semantic evaluation, we adopt standard clustering metrics in plane estimation, including Variation of Information (VOI), Rand Index (RI), and Segmentation Covering (SC). For overall geometry, Chamfer Distance (**Cham.a.**) and F-score are used. We further evaluate the top 20 largest ground-truth planes and report a unified Planar Chamfer metric (**Cham.p.**) for planar geometry. All evaluations follow the protocol of [5].

B. Comparison with baselines

Comparison Results on ScanNetV2. We evaluate on the 100 test splits of ScanNetV2, where PIPS achieves the best performance across all metrics (Table I). M3D+RANSAC [42] benefits from metric prediction but lacks planar awareness, resulting in weak semantics. Airplanes [5] and NeuralPlane [20] incorporate planar semantics yet suffer from weak geometric constraints, leading to distorted planes and irregular segmentation. PlanarSplatting [32] prioritizes geometric fitting while neglecting semantics, causing overall degradation. Qualitative results in Fig. 5 corroborate these findings: prior methods exhibit distortion and fragmented geometry, whereas PIPS consistently recovers coherent planes with aligned semantics.

Comparison Results on ScanNet++. We further evaluate on 30 ScanNet++ scenes, where high-fidelity laser scans and fine-grained annotations make planar reconstruction more challenging, resulting in overall performance drops (Table II). Nevertheless, PIPS achieves the best results across all metrics. Metric3D and SimpleRecon, partly pretrained on ScanNetV2, show limited generalization with large F-score gaps. PlanarSplatting attains comparable geometric accuracy via per-scene optimization but lacks semantic consistency. Airplanes and NeuralPlane also struggle with generalization, producing distorted or irregular planes. Qualitative results in Fig. 5 further confirm these findings: baselines exhibit

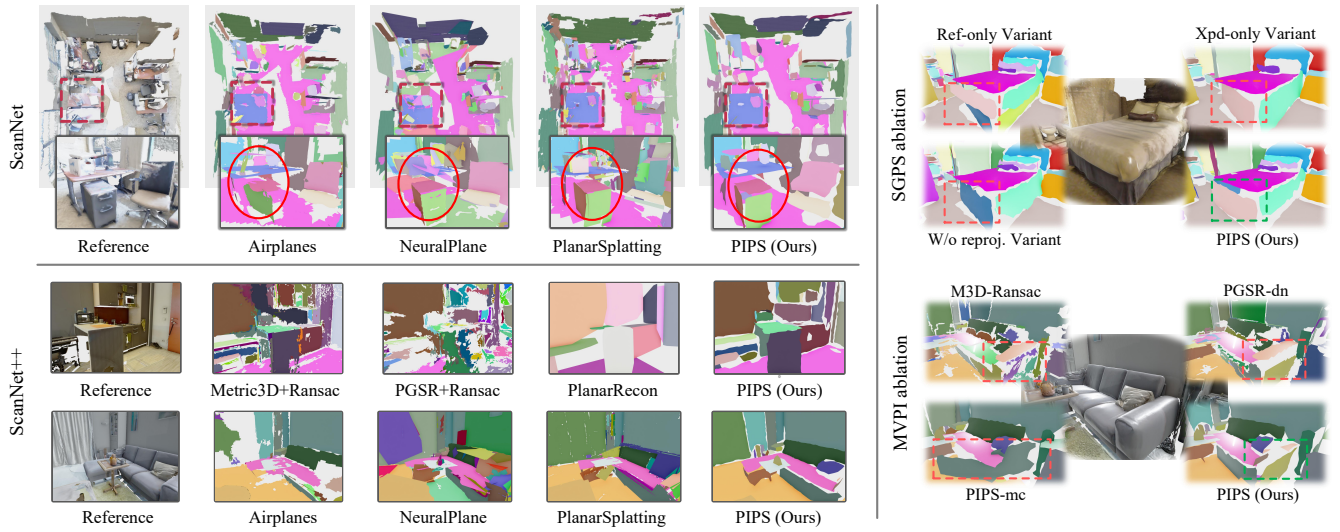


Fig. 5. 3D plane reconstruction results on ScanNetV2 (scene0653) and ScanNet++ (*f38b0108a1*, top; *8a35ef3cfe*, bottom). The bird's-eye view illustrates the abstraction and compression effect of planar reconstruction, while the zoomed-in views highlight semantic and geometric alignment in structural details. Zoomed-in comparisons among SGPS, MVPI ablation variants, and our PIPS further validate the effectiveness of our design.

TABLE III. Ablation results on SGPS (top-three) and MVPI (medium-three). Our PIPS performs the best metrics.

Variants	VOI↓	RI↑	SC↑	Cham.p.↓	Cham.a.↓	F-score↑
Ref-only	2.510	0.965	0.517	11.863	5.197	67.150
XPD-only	2.517	0.963	0.536	10.252	5.414	65.424
w/o reproj	2.570	0.966	0.528	10.663	5.868	63.018
M3D-Ran.	2.597	0.964	0.501	12.165	5.645	63.923
PGSR-dn	2.588	0.963	0.500	13.029	6.138	62.543
PIPS-mc	2.493	0.963	0.535	11.467	5.517	66.559
PIPS (ours)	**2.412**	**0.966**	**0.551**	**9.634**	**5.124**	**68.236**

fragmented or misaligned planes, whereas PIPS reconstructs coherent planar structures with accurate semantic–geometric alignment.

C. Ablation Studies

In this section, we conduct comprehensive ablations on the two key modules of PIPS: (1) the Structure-Guided Planar Segmentor (Sec. III-A) and (2) the Multi-View Planar Integrator (Sec. III-B). All experiments are performed on 6 from ScanNetV2 [46] and 4 from ScanNet++ [44].

SGPS ablation studies. Based on the design of the SGPS module, we evaluate three variants: (1) Ref-only: directly using the refined segments $\mathcal{M}_t^{ref}$ as SGPS segments; (2) Xpd-only: directly using sparse planar segments $\mathcal{M}_t^{xpd}$ as SGPS segments; (3) w/o reproj: directly sampling the SGPS point cloud $\mathcal{P}_I$ from the TSDF-fusion mesh. Quantitative results on 10 ablation scenes are reported in Table III (top). The fusion of $\mathcal{M}_t^{ref}$ and $\mathcal{M}_t^{xpd}$ effectively balances the over-segmentation and under-segmentation drawbacks of the two individual approaches, as shown in Fig. 5 (top-right). Moreover, our proposed meshing–rendering–projection strategy mitigates the planar distortion observed in the w/o reproj variant.

MVPI ablation studies. Based on the design of the MVPI module, we evaluate the following ablation variants: (1) M3D-Ran.: the M3D+Ransac basline; (2) PGSR-dn: PGSR [25] supervised by the same monocular 2.5D cues as PIPS, followed by Seq.Ransac; and (3) PIPS-mc: the MVPI module without planar normal guidance. Quantitative results on 10 ablation scenes are reported in Table III (medium). The superior results indicate that the definition of planar instances in PIPS more effectively exploits 2D structural priors than PGSR-dn and M3D-Ran.. Moreover, PIPS guided by planar normals achieves more accurate planar decomposition than PIPS-mc. The qualitative comparison is shown in Fig. 5 (bottom-right).

V. Conclusion

In this work, we propose **PIPS**, a planar instance 3D reconstruction method that leverages planar structural priors. Starting from planar-unaware geometric predictions, PIPS consistently exploits the prior that "*pixels or points from the same plane share similar normals and distances*" to perform single-view planar segmentation and guide multi-view instance association. Planar semantics and geometry are then bridged at the instance level, with planar structural priors enforcing alignment in both 2D and 3D domains. Extensive experiments on hundreds of indoor scenes demonstrate that PIPS achieves superior planar reconstruction closely matching RGB references. Detailed ablation studies further validate the effectiveness of each designed component.

However, PIPS relies on accurate camera poses, as our current focus is limited to mapping under known trajectories. Since localization drift frequently arises in textureless indoor scenes, as noted in Planar-SLAM [47], extending PIPS towards Planar-based SLAM would be a promising future direction.

References

[1] Z. Jin, T. Tillo, W. Zou, Y. Zhao, and X. Li, "Robust plane detection using depth information from a consumer depth camera," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 447–460, 2017.

[2] Y. Tang, M. Wang, Y. Deng, Z. Zheng, J. Deng, S. Zuo, and Y. Yue, "Openin: Open-vocabulary instance-oriented navigation in dynamic domestic environments," *IEEE Robotics and Automation Letters*, no. 99, pp. 1–8, 2025.

[3] Apple Developer Documentation. Placing content on detected planes. (2024). [Online]. Available: https://developer.apple.com/documentation/visionos/placing-content-on-detected-planes

[4] Google for Developers. Arcore. fundamental concepts: Environmental understanding. (2024). [Online]. Available: https://developers.google.com/ar/develop/fundamentals

[5] J. Watson, F. Aleotti, M. Sayed, Z. Qureshi, O. Mac Aodha, G. Brostow, M. Firman, and S. Vicente, "Airplanes: Accurate plane estimation via 3d-consistent embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5270–5280, 2024.

[6] C. Liu, K. Kim, J. Gu, Y. Furukawa, and J. Kautz, "Planercnn: 3d plane detection and reconstruction from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4450–4459, 2019.

[7] C. Liu, J. Yang, D. Ceylan, E. Yumer, and Y. Furukawa, "Planenet: Piece-wise planar reconstruction from a single rgb image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2579–2588, 2018.

[8] C. D. Duc and J. Lim, "X-pdnet: Accurate joint plane instance segmentation and monocular depth estimation with cross-task distillation and boundary correction," *arXiv preprint arXiv:2309.08424*, 2023.

[9] Y. Qian and Y. Furukawa, "Learning pairwise inter-plane relations for piecewise planar reconstruction," in *European Conference on Computer Vision*, pp. 330–345. Springer, 2020.

[10] C. Sun, C.-W. Hsiao, N.-H. Wang, M. Sun, and H.-T. Chen, "Indoor panorama planar 3d reconstruction via divide and conquer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11 338–11 347, 2021.

[11] J. Shi, S. Zhi, and K. Xu, "Planerectr: Unified query learning for 3d plane recovery from a single view," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9377–9386, 2023.

[12] J. Liu, R. Yu, S. Chen, S. X. Huang, and H. Guo, "Towards in-the-wild 3d plane reconstruction from a single image," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27 027–27 037, 2025.

[13] C. Sommer, Y. Sun, L. Guibas, D. Cremers, and T. Birdal, "From planes to corners: Multi-purpose primitive detection in unorganized 3d point clouds," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1764–1771, 2020.

[14] C. Feng, Y. Taguchi, and V. R. Kamat, "Fast plane extraction in organized point clouds using agglomerative hierarchical clustering," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6218–6225. IEEE, 2014.

[15] A. Roychoudhury, M. Missura, and M. Bennewitz, "Plane segmentation using depth-dependent flood fill," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2210–2216. IEEE, 2021.

[16] M. FISCHLER AND, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[17] M. Sayed, J. Gibson, J. Watson, V. Prisacariu, M. Firman, and C. Godard, "Simplerecon: 3d reconstruction without 3d convolutions," in *European Conference on Computer Vision*, pp. 1–19. Springer, 2022.

[18] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

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

[20] H. Ye, Y. Liu, Y. Liu, and S. Shen, "Neuralplane: Structured 3d reconstruction in planar primitives with neural fields," in *The Thirteenth International Conference on Learning Representations*, 2025.

[21] Z. Chen, Q. Yan, H. Zhan, C. Cai, X. Xu, Y. Huang, W. Wang, Z. Feng, L. Liu, and Y. Xu, "Planarnerf: Online learning of planar primitives with neural radiance fields," *arXiv preprint arXiv:2401.00871*, 2023.

[22] F. G. Zanjani, H. Cai, H. Ackermann, L. Mirvakhabova, and F. Porikli, "Planar gaussian splatting," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8905–8914. IEEE, 2025.

[23] M. Yan, J. Zhang, Y. Zhu, and H. Wang, "Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28 274–28 284, 2024.

[24] F. G. Zanjani, H. Cai, Y. Zhu, L. Mirvakhabova, and F. Porikli, "Neural mesh fusion: Unsupervised 3d planar surface understanding," in *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 353–359. IEEE, 2024.

[25] D. Chen, H. Li, W. Ye, Y. Wang, W. Xie, S. Zhai, N. Wang, H. Liu, H. Bao, and G. Zhang, "Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction," *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[26] W. Zhao, J. Liu, S. Zhang, Y. Li, S. Chen, S. X. Huang, Y.-J. Liu, and H. Guo, "Monoplane: Exploiting monocular geometric cues for generalizable 3d plane reconstruction," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8481–8488. IEEE, 2024.

[27] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[28] S. Agarwala, L. Jin, C. Rockwell, and D. F. Fouhey, "Planeformers: From sparse view planes to 3d reconstruction," in *European Conference on Computer Vision*, pp. 192–209. Springer, 2022.

[29] L. Jin, S. Qian, A. Owens, and D. F. Fouhey, "Planar surface reconstruction from sparse views," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12 991–13 000, 2021.

[30] B. Tan, N. Xue, T. Wu, and G.-S. Xia, "Nope-sac: Neural one-plane ransac for sparse-view planar 3d reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15 233–15 248, 2023.

[31] Y. Xie, M. Gadelha, F. Yang, X. Zhou, and H. Jiang, "Planarrecon: Real-time 3d plane detection and reconstruction from posed monocular videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6219–6228, 2022.

[32] B. Tan, R. Yu, Y. Shen, and N. Xue, "Planarsplatting: Accurate planar surface reconstruction in 3 minutes," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1190–1199, 2025.

[33] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

[34] Y. Deng, J. Wang, J. Zhao, X. Tian, G. Chen, Y. Yang, and Y. Yue, "Opengraph: Open-vocabulary hierarchical 3d graph representation in large-scale outdoor environments," *IEEE Robotics and Automation Letters*, vol. 9, no. 10, pp. 8402–8409, 2024.

[35] Y. Deng, Y. Yue, J. Dou, J. Zhao, J. Wang, Y. Tang, Y. Yang, and M. Fu, "Omnimap: A general mapping framework integrating optics, geometry, and semantics," *IEEE Transactions on Robotics*, 2025.

[36] R. Huang, S. Peng, A. Takmaz, F. Tombari, M. Pollefeys, S. Song, G. Huang, and F. Engelmann, "Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels," in *European Conference on Computer Vision*, pp. 278–295. Springer, 2024.

[37] P. Nguyen, T. D. Ngo, E. Kalogerakis, C. Gan, A. Tran, C. Pham, and K. Nguyen, "Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4018–4028, 2024.

[38] Y. Yang, X. Wu, T. He, H. Zhao, and X. Liu, "Sam3d: Segment anything in 3d scenes," *arXiv preprint arXiv:2306.03908*, 2023.

[39] L. Qi, J. Kuen, T. Shen, J. Gu, W. Li, W. Guo, J. Jia, Z. Lin, and M.-H. Yang, "High quality entity segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4047–4056, 2023.

[40] Y. Deng, J. Wang, J. Zhao, J. Dou, Y. Yang, and Y. Yue, "Openobj: Open-vocabulary object-level neural radiance fields with fine-grained understanding," *IEEE Robotics and Automation Letters*, 2024.

[41] T. Pan, L. Tang, X. Wang, and S. Shan, "Tokenize anything via prompting," in *European Conference on Computer Vision*, pp. 330–348. Springer, 2024.

[42] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen, "Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal

estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[43] K. C. Gowda and G. Krishna, "Agglomerative clustering using the concept of mutual nearest neighbourhood," *Pattern recognition*, vol. 10, no. 2, pp. 105–112, 1978.

[44] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai, "Scannet++: A high-fidelity dataset of 3d indoor scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

[45] F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, and G. Taubin, "The ball-pivoting algorithm for surface reconstruction," *IEEE transactions on visualization and computer graphics*, vol. 5, no. 4, pp. 349–359, 2002.

[46] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

[47] Y. Li, R. Yunus, N. Brasch, N. Navab, and F. Tombari, "Rgb-d slam with structural regularities," in *2021 IEEE international conference on Robotics and automation (ICRA)*, pp. 11 581–11 587. IEEE, 2021.