Planar Gaussian Splatting

Farhad G. Zanjani Hong Cai Hanno Ackermann Leila Mirvakhabova Fatih Porikli

Qualcomm AI Research*

{fzanjani, hongcai, hackerma, lmirvakh, fporikli}@qti.qualcomm.com

Abstract

This paper presents Planar Gaussian Splatting (PGS), a novel neural rendering approach to learn the 3D geometry and parse the 3D planes of a scene, directly from multiple RGB images. The PGS leverages Gaussian primitives to model the scene and employ a hierarchical Gaussian mixture approach to group them. Similar Gaussians are progressively merged probabilistically in the tree-structured Gaussian mixtures to identify distinct 3D plane instances and form the overall 3D scene geometry. In order to enable the grouping, the Gaussian primitives contain additional parameters, such as plane descriptors derived by lifting 2D masks from a general 2D segmentation model and surface normals. Experiments show that the proposed PGS achieves state-of-the-art performance in 3D planar reconstruction without requiring either 3D plane labels or depth supervision. In contrast to existing supervised methods that have limited generalizability and struggle under domain shift, PGS maintains its performance across datasets thanks to its neural rendering and scene-specific optimization mechanism, while also being significantly faster than existing optimization-based approaches.

1. Introduction

Identifying 3D planar surfaces in indoor settings using multi-view posed monocular video is a pre-requisite for many applications, including augmented reality, virtual reality, robot navigation, and 3D interior modeling. Since man-made environments feature many diverse planar surfaces whose appearances can be ambiguous, this is a challenging task. By approximating scene geometry with a collection of basic planar shapes, we achieve a compact and efficient representation that facilitates interaction with the physical space.

Recent deep learning methods treat 3D planar surface understanding as supervised learning tasks, relying on annotations of either 2D planes [1, 19, 20, 34, 45] or 3D structures [40]. However, acquiring plane annotations in both high-quality and large-scale is an expensive endeavor. Furthermore, these models struggle to generalize to unseen scenes or those captured with different imaging sensors.

Recent advancements in differentiable rendering enable 3D geometry reconstruction solely from multi-view 2D images, eliminating the need for 3D ground truth. While methods like Neural Radiance Fields (NeRF) and their successors [7, 24, 36] achieve impressive novel view synthesis (NVS) quality, it remains a challenge to extract explicit planar surfaces from their implicit representations [2]. Specifically, volume-based approaches [38, 42, 44] rely on computationally expensive steps like ray marching and density field prediction for implicit surface modeling, followed by Marching Cubes [23] for surface extraction and Sequential RANSAC [10] for plane detection. These steps require careful tuning of numerous hyperparameters (e.g., in RANSAC), adding complexity and hindering broader application.

Comparing to implicit methods, explicit neural representations offer several advantages. They allow direct optimization of the geometry through volumetric tetrahedral mesh [5, 12, 26, 30], triangle surface mesh [46], or point cloud [16] on the geometric primitives themselves. This makes it easier to constrain the reconstructed surfaces, for example, to be locally planar. However, most of existing explicit methods are primarily developed for novel view synthesis and require additional steps for planar reconstruction. Among existing explicit neural approaches, recently NMF [46] has proposed direct optimization on the 3D vertex positions of a triangle mesh to jointly reconstruct the geometry and perform contrastive learning for 3D planar parsing.

In this paper, we propose, Planar Gaussian Splatting (PGS), to represent planar surfaces with a set of Gaussian primitives, equipped with learned plane descriptors, which are jointly optimized with other Gaussian parameters, *i.e.*, without requiring complex and error-prone post-hoc heuristics. More specifically, we propose a hierarchical, tree-

*Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

structured Gaussian Mixture Model (GMM) to model the scene. This probabilistic approach permits grouping the Gaussian geometric primitives trained by Gaussian Splatting while providing an interpretable interface for parsing and optimizing for planes. In order to facilitate the grouping, we leverage 2D segmentation from foundation models like SAM [17]. This allows our algorithm to optimize multi-view partial consistency between 2D segmentation pseudo labels to identify 3D planes while optimizing for the 3D geometry concurrently.

Our main contributions are summarized as follows:

- We propose Planar Gaussian Splatting (PGS), a novel unsupervised, neural-rendering-based framework for 3D planar reconstruction of a scene from RGB images. PGS does not require any 3D labels or depth supervision (either ground-truth or predicted).
- We leverage 3D Gaussian primitives to model the scene, which are probabilistically grouped via hierarchical, tree-structured Gaussian mixtures to identify 3D planar instances.
- In order to enable the grouping, we propose discriminative 3D descriptors, learned from Segment Anything [17] proposal masks. We resolve challenges such as the lack of multi-view proposal association and variable numbers of masks, by formulating and solving a linear regression problem followed by merging segments using a Region Adjacency Graph (RAG).
- PGS achieves state-of-the-art performance in 3D planar reconstruction, as compared to existing supervised and optimization-based methods. In particular, it can readily reconstruct 3D planes on any new test scenes, which cannot be done well by supervised learning models.

2. Related work

2.1. Planar Reconstruction

Planar reconstruction from a single RGB image have been investigated in several works, for instance using ConvNets [20, 41]. Those works predict both segmentation and 3D plane parameters and furthermore require a prescribed maximum number of planes in an image, which limits model applicability. Other works address this limitation on single image planar reconstruction [19,28,45] and can handle any number of planes. PlaneTR [34] leverages transformers to consider context information and geometric cues like line segmentation and ground-truth depth in a sequence-to-sequence way. Dependencies on ground-truth depth or plane annotation and single image reconstruction limit the applications of these approaches.

Alternatively, multi-view reconstruction utilizes multiple images, which contain richer geometric information. Several works share a common two-stage approach: local plane detection and plane parameter estimation [15, 21]. More recently, PlanarRecon [40] proposes to detect planes from video fragments and combine them to create a comprehensive global planar reconstruction, which is supervised by 3D ground-truth planes in training. In contrast, this paper presents a multi-view 3D planar surface reconstruction method without requiring 2D or 3D plane annotations.

2.2. Objects & Semantics in Volumetric Rendering

In recent years, there has been significant progress in radiance field rendering, particularly in the context of semantic 3D modeling and scene decomposition.

Neural scene representations for scene decomposition: Some works [9,25,29,35,39,43] employ neural scene representations to decompose scenes into foreground and background components, remarkably without explicit supervision or only relying on weak signals such as text or object motion. Semantic NeRF [47], for instance, introduces a separate branch that predicts semantic labels. NeSF [37] predicts a semantic field by utilizing a density field as input to a 3D semantic segmentation model.

Supplementing NeRFs with 2D annotations: Some works have explored ways to enhance NeRF models using readily available 2D annotations from datasets. Panoptic NeRF [11] and Instance-NeRF [22] incorporate 3D instance supervision, allowing for more accurate scene representation. Contrastive Lift [3] takes a novel approach by lifting 2D instance segmentation to 3D without relying on explicit 3D masks. It achieves this through contrastive learning. Similarly, NMF [46] introduces a contrastive learning scheme for lifting 2D pixel clusters into a 3D mesh. Panoptic Lifting [31] addresses the challenge of lifting 2D instance segmentation by employing linear assignment techniques to ensure consistency across multi-view annotations.

Fusing 2D analysis output into 3D space: The aforementioned works — Semantic NeRF [47], Panoptic Lifting [31] and Constrastive Lift [3] — represent a recent research direction. They aim to seamlessly integrate the insights gained from 2D analysis into the 3D domain. SA3D [4], leverages the powerful vision foundation model SAM [17]. SAM excels at segmenting objects in 2D images, and SA3D utilizes this capability for interactive 3D segmentation using NeRF.

Our work shares similar concept with the above mentioned method since the proposed method lifts the 2D analysis into the 3D domain but has a different objective which is 3D planar reconstruction of the scene. A more recent and most similar work, NMF [46] introduces an explicit rendering approach for 3D planar reconstruction by optimizing the vertices of triangle mesh and incorporating a contrastive learning for jointly learning the scene geometry and decomposition the mesh into planar surfaces. NMF leverages a depth prediction network for lifting the 2D analysis into 3D. Our approach doesn't require the ground truth

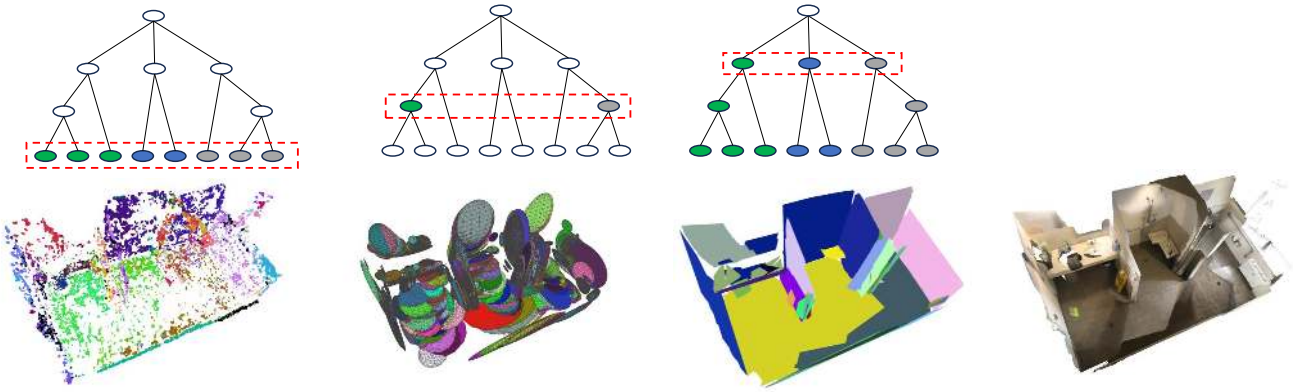


Figure 1. The proposed PGS method constructs the entire scene using a tree-structured arrangement of Gaussian nodes. At the leaf nodes, Gaussian primitives are optimized through Gaussian splatting rendering. Meanwhile, parent nodes are recursively formed by merging Gaussian nodes probabilistically. The child nodes of the root correspond to distinct 3D planes within the scene, as depicted in the accompanying figure.

or predicted depth network. Furthermore, similar to Constrastive Lift [3], the NMF utilizes a clustering method as post-processing for grouping the features of the plane or object instances. The constant hyper-parameters of clustering introduces a sub-optimal solution across different scene, while our approach utilizes a probabilistic method for grouping the features.

These advancements bridge the gap between 2D and 3D representations, opening up exciting possibilities for richer scene understanding and modeling.

3. Planar Gaussian Splatting

In this section, we present the proposed Planar Gaussian Splatting (PGS) method. Section 3.1 starts with a primer on Gaussian Splatting and the standard parameterization of the Gaussian primitives. Section 3.2 introduces the Gaussian Mixture Tree, a probabilistic approach for planar construction of the scene geometry in a bottom-up modeling. Section 3.3 introduces a learning procedure for optimizing a latent vector per Gaussian primitives, called plane descriptor which represents 3D plane instances. Afterwards, local planar alignment as a geometric constraint applies on Gaussian positions close to surfaces is explained in Section 3.4. In Section 3.5, we discussed how a holistic separability across descriptors is maintained using recurrent mean-shift layer.

3.1. Primer on 3D Gaussian Splatting (3DGS)

3DGS [16] models the scene as a set of multivariate Gaussians in 3D space, which is an explicit form of representation, in contrast to the implicit representation used in NeRF. Each Gaussian is characterized by a covariance matrix Σ and a center (mean) point μ, *i.e.*,

$$G(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}. \tag{1}$$

The centers of these 3D Gaussians are initialized from a set of sparse points (*e.g.*, randomly initialized or obtained from SfM). More specifically, each Gaussian is parameterized by the following parameters: (a) a center position $\mu \in \mathbb{R}^3$, (b) a covariance matrix which has the form $\Sigma = RSS^TR^T$ computed from scaling $\mathbf{s} \in \mathbb{R}^3$ and rotation factors $\mathbf{r} \in \mathbb{R}^4$ (in quaternion), (c) opacity $\alpha \in \mathbb{R}$, and (d) spherical harmonics (SH) coefficient $\mathbf{c} \in \mathbb{R}^k$ that represents the color, where k is the degree of the SH. Given a view transform W, the 2D covariance matrix in camera coordinates can be expressed by $\Sigma^{2D} = JW\Sigma W^TJ^T$, where J is the Jacobian of the affine approximation of the projection transformation. For each pixel in the image, rendering the color is performed as in [16] by blending the color vectors of N ordered Gaussians, which overlap at the pixel position (u, v), by

$$\begin{aligned} \hat{\mathbf{c}}_{uv} &= \sum_{i}^{N} \mathbf{c}_i \alpha_i \prod_{j}^{i-1}(1-\alpha_j), \\ \mathcal{L}_{\text{rgb}} &= \sum_{(u,v)} \|\hat{\mathbf{c}}_{uv} - \mathbf{c}_{uv}\|_1 + \lambda \cdot \texttt{SSIM}(\hat{\mathbf{c}}_{uv}, \mathbf{c}_{uv}), \end{aligned} \tag{2}$$

where α_i and $\mathbf{c}_i$ are learnable opacity and RGB color of i^{th} Gaussian, obtained by SH coefficients. The weighting coefficient λ is set the same as in [16]. By minimizing the loss in Eq. 2, the model learns the 3D scene geometry through the sparse unstructured Gaussian primitives by optimizing their opacity parameters. In addition to optimizing the Gaussian parameters, the training process involves splitting, cloning, and culling of Gaussian primitives to express the scene geometry by maximizing the photometric likelihood between the rendered and the actual images, as proposed in [16].

3.2. Gaussian Mixture Tree

Optimizing the parameters of 3D Gaussians leads to spatially moving the center points close to the object surfaces in

the scene. In order to identify distinct 3D plane instances, a novel probabilistic approach is proposed that involves compositional modeling of the 3D scene using a Gaussian Mixture Tree (GMT).

The whole scene is modeled in a tree structure, which is constructed recursively from the leaf nodes to the root node. The Gaussian Mixture Model (GMM) is involved to join nodes and derive intermediate parent nodes. In this way, the GMT represents the entire scene in a hierarchical way. In the GMT, the child nodes of the root represent individual 3D plane instances; see Figure 1.

Each parent in the tree (except the root) specify a Gaussian distribution, $G_p(\mu_p, \Sigma_p)$, in 3D space, which encompasses all the center points (μ) of its child Gaussian nodes. Since each node specifies a Gaussian distribution, merging two nodes is equivalent to merging their distributions. More specifically, for two Gaussian nodes, the merging is performed as follows.

$$\begin{aligned}\mathbf{\Sigma}_p =& \mathbf{\Sigma}_j \cdot (\mathbf{\Sigma}_i + \mathbf{\Sigma}_j)^{-1} \cdot \mathbf{\Sigma}_i, \\ \mu_p =& \mathbf{\Sigma}_j \cdot (\mathbf{\Sigma}_i + \mathbf{\Sigma}_j)^{-1} \cdot \mu_i + \mathbf{\Sigma}_i \cdot (\mathbf{\Sigma}_i + \mathbf{\Sigma}_j)^{-1} \cdot \mu_j,\end{aligned} \tag{3}$$

where $(\mu_p, \mathbf{\Sigma}_p)$ specifies the Gaussian parameters of the parent node.

The merging criteria are based on both the Bhattacharya distance [14] between the two respective Gaussian distributions, and the cosine similarity between the descriptors of the nodes $\langle \mathbf{z}_i \cdot \mathbf{z}_j \rangle$ (which will be explained in the following part). For two multivariate Gaussian distributions (G_i, G_j), the Bhattacharya distance $D_B(G_i, G_j)$ is given by:

$$\begin{aligned}D_B(G_i, G_j) =& \frac{1}{8}(\mu_i - \mu_j)^T \mathbf{\Sigma}^{-1}(\mu_i - \mu_j) \\ &+ \frac{1}{2}\ln\left(\frac{\det \mathbf{\Sigma}}{\sqrt{\det \mathbf{\Sigma}_i \det \mathbf{\Sigma}_j}}\right),\end{aligned} \tag{4}$$

where $\mathbf{\Sigma} = \frac{\mathbf{\Sigma_i} + \mathbf{\Sigma_j}}{2}$. The tree structure simply is formed by merging every two nodes whose descriptors are similar and whose Bhattacharya distance is lower than a predefined threshold. Algorithm 1 shows pseudocode of the proposed GMT.

When constructing the GMT, we do not directly use the Gaussian primitives (*i.e.*, parameters obtained through Gaussian splatting optimization) at the leaf nodes, which would be computationally expensive given as there can be a huge number of them in the 3D scene. Instead, we first group the Gaussian primitives into local clusters and use each cluster as the leaf node of the GMT. For locally grouping the Gaussian primitives, two additional parameters are introduced for each Gaussian during the optimization, which are the surface normal vector at the center location of the Gaussian and a learnable vector called *plane descriptor* which can be used to identify distinct 3D planes.

Algorithm 1 Hierarchically Merging Gaussians

Input : $\{G_i(\mu_i, \mathbf{\Sigma}_i, \mathbf{n}_i, \mathbf{z}_i)\}_{i=1}^{N_s}$, and N_s leaf nodes
Output: $p(\mathbf{x}|\mathbf{G}^s) = \sum_{k=1}^{L} \pi_k \cdot p(\mathbf{x}|\mathbf{G}_k^s)$, where $x \in \mathbb{R}^3$
Initialize: $L = 0, G^s = \emptyset$
for *i=1:*N_s **do**
  if *i is NOT descendant of* $\{G_k^s\}_{k=1}^{L}$ **then**
    for *j=i+1:*N_s **do**
      if *j is NOT descendant of* $\{G_k^s\}_{k=1}^{L}$ **then**
        if $D_B(G_i, G_j) \leq \epsilon_B$ *and* $1 - \langle \mathbf{z}_i \cdot \mathbf{z}_j \rangle \leq \epsilon_z$ **then**
          $G_p^{ij}(\mu_p, \mathbf{\Sigma}_p) \leftarrow G_i + G_j$ // (3)
          $G_i \leftarrow G_p^{ij}$
        end
      end
    end
    $G^s.insert(G_i)$ // Add to plane nodes
    $L \leftarrow L + 1$
  end
end

3.3. Learning Plane Descriptors

To organize Gaussian primitives within the GMT hierarchy, two additional parameters for each Gaussian primitive are introduced: a normal vector $n \in \mathbb{R}^3$ and a plane descriptor $z \in \mathbb{R}^k$ (*e.g.*, $k = 3$).

Lifting 2D normal maps to 3D: To learn the normal vectors in 3D field, we employ an off-the-shelf network that predicts the normal map for the 2D training images; specifically, we use the Omnidata model [8]. To lift the normal vectors from 2D to 3D, similar to Eq. 2, the normal vectors are rendered for the camera view and compared with the off-the-shelf network's prediction $(\mathbf{n}_{uv})$ at pixel position (u, v), using the cosine distance. The normal loss is defined as:

$$\mathcal{L}_{\text{n}} = \sum_{(u,v)} \left(1 - \langle \hat{\mathbf{n}}_{uv} \cdot \mathbf{n}_{uv} \rangle\right), \ \hat{\mathbf{n}}_{uv} = \sum_{i}^{N} \mathbf{n}_i \alpha_i \prod_{j}^{i-1}(1 - \alpha_j). \tag{5}$$

Lifting 2D SAM masks to 3D: To learn the plane descriptors, we leverage the 2D masks generated by the Segment Anything Model (SAM) [17]. SAM segments the input 2D images into object parts. We prompt SAM with 32×32 regular grid points. For each point, SAM predicts a set of masks that may correspond to different parts of objects. SAM incorporates ambiguity-aware modeling: if a point lies on a part or subpart, SAM may return the subpart, the part, or the entire object. The image segments with high variance in their normal vectors are ignored and are considered as invalid regions for the current camera frame.

In order to learn the plane descriptors in the 3D Gaussian field, the valid 2D image segments need to be lifted to 3D. However, the masks from SAM have no semantic information and 2D-3D lifting is not straightforward due to two challenges: (1) the mask associations across different views are unknown, and (2) the number of segments is variable in

each image and the maximum number of segments is unknown. In the following, we discuss how we handle these issues and utilize the 2D segments to learn the 3D plane descriptors.

We restrict the descriptor $\mathbf{z}$ to have a vector norm equal to one ($\|\mathbf{z}\|_2 = 1$). To address the aforementioned challenge of lifting the 2D segments to create 3D descriptors, we propose a linear regression approach (with closed-form solution) to predict the indices of 2D segments based on $\mathbf{z}$, for each individual training image. More specifically, given the camera pose, we first render a 2D descriptor image where each pixel is computed by blending the 3D descriptors of Gaussian field for given camera view. Next, we use a linear layer ($\mathbf{W}$) to map each pixel in this descriptor image to a one-hot vector $\mathbf{y}$ that encodes a segment. $\mathbf{W}$ can be solved analytically. In matrix form, this is specified as follows.

$$\begin{aligned} \mathbf{Y} = [\mathbf{Z}|\mathbf{1}]\cdot\mathbf{W} \;\rightarrow\; \hat{\mathbf{W}} = (\mathbf{Z}^T\cdot\mathbf{Z})^{-1}\cdot\mathbf{Z}^T\cdot\mathbf{Y}, \\ \mathcal{L}_{seg} = \sum_i \|y_i - \hat{y}_i\|_1, \end{aligned} \tag{6}$$

where $\mathbf{y}_i \in \{0,1\}^m$ and $\sum_{j=1}^{m} y_{ij} = 1$. The $\mathbf{Y}$ and $\mathbf{Z}$ denote the matrix form of the labels and descriptors, the loss is computed by comparing the prediction $\hat{\mathbf{y}} = [\mathbf{z}|\mathbf{1}]\cdot\hat{\mathbf{w}}$ with the labels $\mathbf{y}$.

Eq. 6 optimizes the descriptors $\mathbf{z}$ for the Gaussians, taking advantage of the fast rendering of 3DGS. Note that the linear regression of Eq. 6 is recomputed for each given camera view and the length m of the target vector $\mathbf{y}_i$ can be variable in each image, depending on the number of segments in $\mathbf{Y}$. Figure 2 provides visual examples of the learned descriptors and SAM segmentation masks.

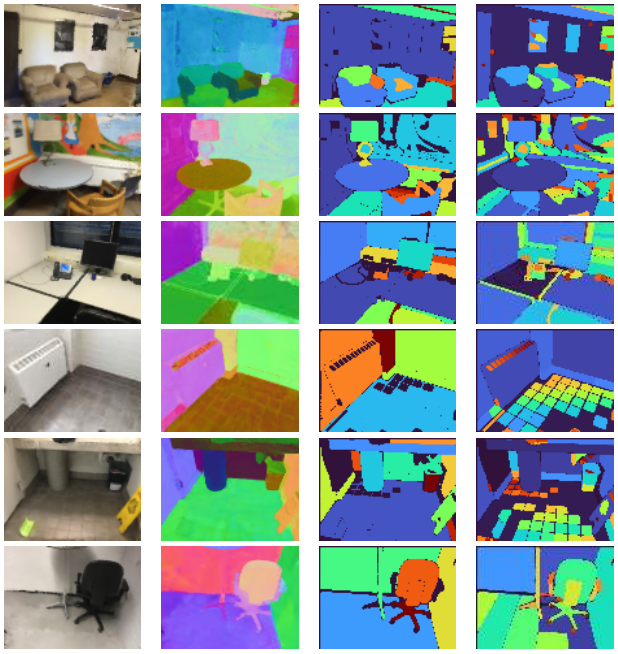
Since the descriptors need to represent distinct 3D planes of the scene, we cannot directly use the SAM segments as the label $\mathbf{Y}$ as there can be multiple segments on the same plane. 2D segments belonging to the same planar surface should be merged. Since we do not have access to 3D or 2D plane annotations, we perform this merging using a **Region Adjacency Graph (RAG)**. The nodes of RAG represent the segments given by SAM and the edges connect nodes whose corresponding regions are adjacent in the image. We partition the RAG by cutting the edges that connect nodes from two different planes and keep the ones from the same plane connected. In order to do that, we use the normal vector of each node as well as a *planar distance*.

The normals can be obtained by rendering as specified in Eq. 5. However, partitioning the RAG solely based on the dissimilarity of normal vectors between neighboring nodes is insufficient; two nodes with similar normals may actually belong to two different planes, *e.g.*, two planes at different heights in the scene. To resolve this ambiguity, we additionally assign a planar distance, d_p, to each node of the RAG. To compute the planar distance for each segment, we assume every pixel belongs to one planar surface in the scene. The corresponding 3D point $\mathbf{p} \in \mathbb{R}^3$ and the plane satisfy the point-normal equation: $\mathbf{n}\cdot\mathbf{p} + d_p = 0$, where $\mathbf{n} = (n_1, n_2, n_3)^\top$ is the normal vector. As such, planar distance d_p at pixel position (u, v) in the image is obtained as follows:

$$d_p = d_{(u,v)} \cdot \left(\frac{n_1}{f_x}(u_0 - u) + \frac{n_2}{f_y}(v_0 - v) - n_3\right), \tag{7}$$

where (f_x, f_y) are the x and y focal lengths of the pinhole camera, and (u_0, v_0) is the principal point, and the depth $d_{(u,v)}$ is computed by rendering the 3D Gaussian field.

The average normal and planar distance values within each SAM segment is assigned to the corresponding node in the RAG. Finally, by thresholding these values, we can cut the edges accordingly in the RAG. Figure 2 (c) shows visual examples of merging SAM segments by partitioning the RAG.



(a) (b) (c) (d)

Figure 2. Visualizing the steps of generating plane descriptors: (a) rendered color image, (b) rendered plane descriptors from 3D Gaussian field in camera view, (c) merged SAM masks by partitioning the Region Adjacency Graph, (d) masks by SAM.

3.4. Local Planar Alignment of 3DGS

The original 3DGS method [16] excels in photo-realistic novel view synthesis through volume rendering and alpha-blending technique. However, it operates without any explicit constraints on the spatial arrangement of the Gaussian primitives, which makes it not directly applicable for learning geometric features. More specifically, the alpha-blending integration of Gaussians (as described in Eq. 2), which are depth-sorted relative to the camera, results in applying the 2D supervision to the overall integration of Gaus-

sians, rather than to each instance individually. For example, occluded Gaussians receive weak supervision, leading to suboptimal parameter optimization. Although this is not critical for rendering RGB images, it poses challenges to learning 3D normal vectors and surface descriptors, which are additional parameters in our proposed PGS. To address this, we enforce the centers of Gaussians to lie exactly on the surfaces of objects. This can be achieved through the alignment process which involves projecting the centers of Gaussian on their local tangent planes. This requires first computing the K Nearest Neighbors (KNN) of Gaussians and then computing the covariance matrices of KNN Gaussian centers in the scene. Given the local covariance 3×3 matrices, the local tangent planes are specified by the two eigenvectors, corresponding to the two largest eigenvalues, of covariance matrices, using singular value decomposition. Estimating the KNN indices also allows us to apply a Laplacian smoothing on the learnt normal and descriptor features by averaging over the features of neighbouring Gaussian primitives.

3.5. Holistic Separability of Gaussian Descriptors

The minimization of the segmentation loss term in Eq. 6 results in a discriminative representation of descriptors denoted as $\mathbf{z}$ in the current image. This representation helps identify Gaussians that belong to distinct 3D plane instances. In order to maintain a holistic separability of descriptors across all planar surfaces in the scene (including surfaces that have never been seen jointly in any camera view), a recurrent mean-shift update [18] is applied to the entire Gaussian field, the matrix form of which is given as follows.

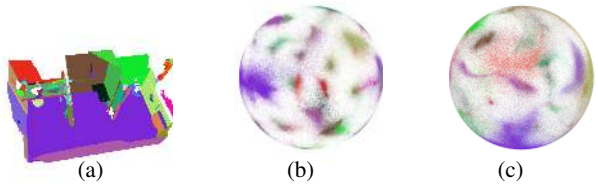
$$\mathbf{Z} \leftarrow \mathbf{Z} \cdot (\eta \cdot \mathbf{K} \cdot \mathbf{D}^{-1} + (1 - \eta) \cdot \mathbf{I}), \tag{8}$$

where $\mathbf{K} = e^{(\gamma \cdot \mathbf{Z}^T \cdot \mathbf{Z})}$ is von Mises-Fisher (vMF) distribution of $\mathbf{z}$ on sphere and $\mathbf{D} = diag(\mathbf{K}^T \cdot \mathbf{I})$ is the diagonal matrix. η is the rate of update and γ is the kernel bandwidth which determines the smoothness of the kernel density estimation.

By applying such updates, we improve separability of the descriptors of all the 3D plane instances. Figure 3 visualize the impact of applying Eq. 8 in the training of PGS. In practice we run updates on $\mathbf{z}$ as specified in Eq. 8 for a few steps at every N iterations in the optimization process. Since the number of Gaussians in the scene can be very large, we use an efficient way to compute Eq. 8; more details on this are provided in the Appendix.

4. Experiments

We conduct experiments to evaluate the 3D plane instance segmentation of PGS, as well as compare with existing competitive approaches. We further perform ablation study to analyze various design choices in the proposed approach.



(a) (b) (c)

Figure 3. Effect of holistic separation. (a) Ground-truth plane labels of an example scene from ScanNet dataset, (b) learned descriptors with applying the holistic separability which results in a more compact and separable representation, (c) learned descriptors without the holistic constraint.

Table 1. 3D plane instance segmentation results on ScanNet. The symbol ++ indicates that the method only produces a 3D mesh reconstruction and a post-processing by Sequential RANSAC is performed to extract the 3D planes.

Method	VOI ↓	RI ↑	SC↑	*Supervision* RGB	Geo.	Planes	Inference time
NeuralRecon++ [33]	5.540	0.696	0.139	✓	✓	-	2 min
PlanarRecon [40]	3.458	0.861	0.359	✓	✓	✓	realtime
3DGS++ [16]	5.056	0.850	0.306	✓	-	-	16 min.
NMF [46]	3.253	0.880	0.381	✓	-	-	40 min.
Planar GS (ours)	**3.045**	**0.901**	**0.430**	✓	-	-	16 min.

Datasets. We perform evaluation on 3D planar reconstruction benchmarks commonly used by prior works [40, 46], including ScanNetv2 [6] and Replica [32]. ScanNetv2 contains RGB videos taken by a mobile device from indoor scenes with the camera pose information associated with each frame. We run our experiments on 10 scenes; 4 of them are the same as in [13]. Replica is a synthetic dataset featuring a diverse set of indoor scenes. Each scene is equipped with high-quality geometry and photo-realistic textures, allowing one to render high-fidelity images from arbitrary camera poses.

Baselines. There are only a few existing works that focus on learning-based multi-view 3D planar reconstruction. We compare PGS with two types of approaches: (1) specialized 3D planar reconstruction methods, *e.g.*, PlanarRecon [40], which is trained with 3D geometry and 3D plane supervisions, and NMF [46] which is an optimization-based approach using depth and normal supervision, (2) dense 3D reconstruction methods like NeuralRecon [33] with 3D geometry supervision and 3DGS [16] with 2D RGB supervision, followed by Sequential RANSAC to extract planes [10]. We denote them as NeuralRecon++ and 3DGS++.

Metrics. Similar to prior works [19, 34, 40, 46], we evaluate the performance of 3D plane instance segmentation by measuring the Rand Index (RI), Variation of Information (VOI), and Segmentation Covering (SC).

4.1. Main Results

On ScanNet: Table 1 shows 3D planar segmentation results on ScanNet data. We see that our proposed PGS achieves significantly better performance across all the eval-

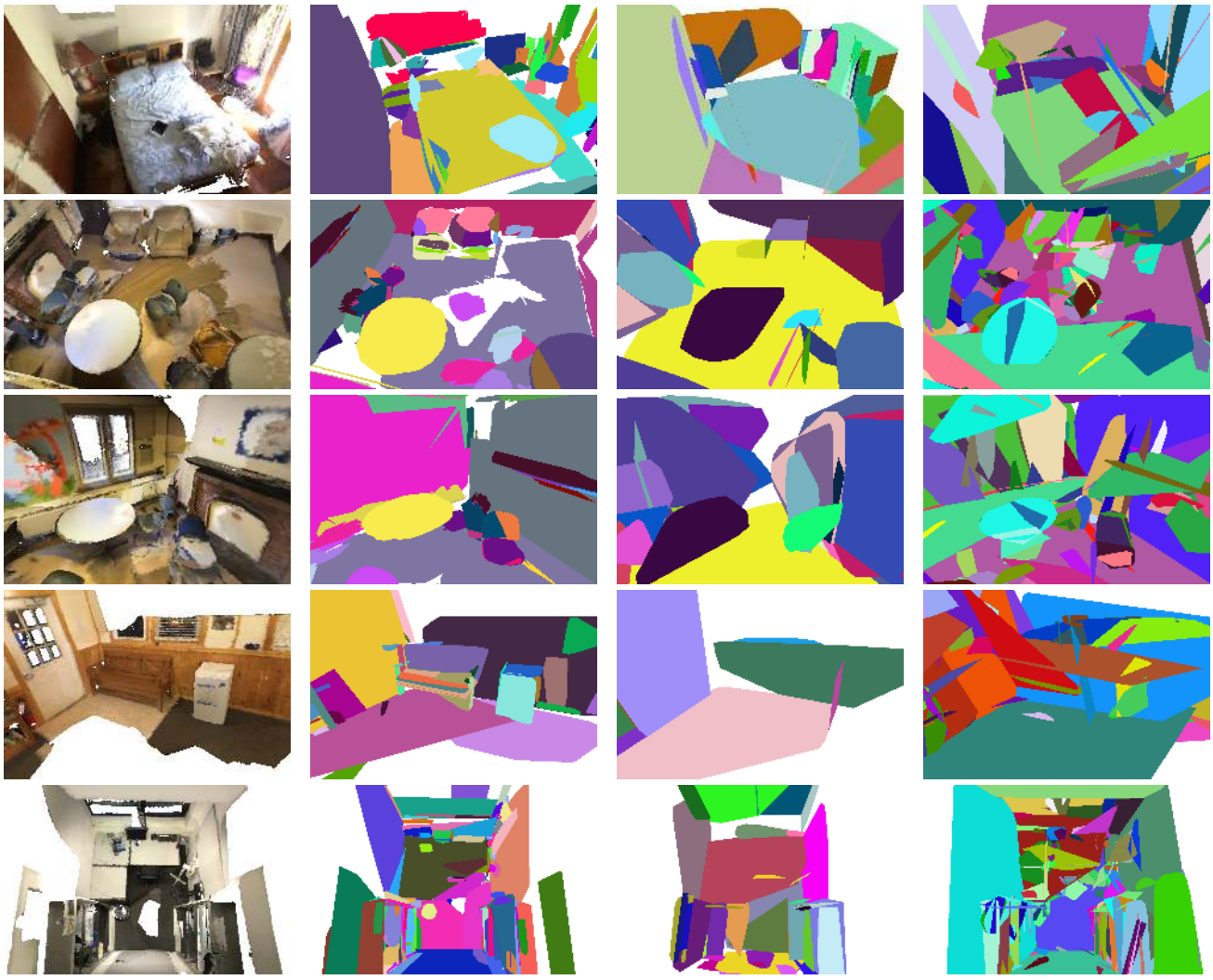


Figure 4. Examples of 3D planar reconstruction on ScanNet. (a) Ground-truth textured meshes (with holes on floor due to unseen regions in the videos), (b) our proposed Planar Gaussian Splatting (PGS), (c) PlanarRecon [40], and (d) 3DGS++. We see that our proposed PGS produces more accurate 3D planes, in terms of precise shapes and recall. PlanarRecon misses many planes and produces incorrect shapes, and 3DGS++ results are noisy and inaccurate.

uation metrics. For instance, it has significant higher segmentation covering score as compared to existing methods (more than 11% improvement). Although both NeuralRecon and PlanarRecon are trained on ScanNet, their 3D plane instance segmentation scores are considerably worse. We also see that naively using Sequential RANSAC to extract 3D planes from 3DGS reconstruction does not yield good accuracy. In terms of runtime, our proposed PGS is significantly faster as compared to the latest state-of-the-art optimization-based 3D plane segmentation method of NMF, with more than 60% less runtime.

Figure 4 shows sample 3D planar reconstruction results on ScanNet. It can be seen that our proposed PGS generates more accurate planes. For instance, in the 2nd row, our PGS captures the shape of the round table top (b) while PlanarRecon fails to recover the shape (c). In addition, PGS has higher recalls. We see that PlanarRecon misses a lot of planes in the scene, especially for smaller objects like chairs, while PGS better identifies planes on those objects. However, the PGS also makes mistake on small subset of the estimated 3D planes by duplicating plane instances, which can be observed at the intersections of planes in our visualizations. The results by 3DGS++ are very noisy, showing that it is suboptimal to simply add a post-processing step like Sequential RANSAC to extract 3D planes from a reconstruction, as compared to an end-to-end, holistically designed pipeline.

On Replica: Table 2 shows the evaluation results on

Table 2. 3D plane instance segmentation results on Replica. The symbol ++ indicates that the method only produces a 3D mesh reconstruction and a post-processing by Sequential RANSAC is performed to extract the 3D planes. Note that ScanNet-trained NeuralRecon fails to produce valid meshes on Replica

Method	VOI ↓	RI ↑	SC↑	*Supervision*		
				RGB	Geo.	Planes
NeuralRecon++ [33]	-	-	-	✓	✓	-
PlanarRecon [40]	4.676	0.829	0.148	✓	✓	✓
3DGS++ [16]	4.401	0.904	0.179	✓	-	-
NMF [46]	4.311	0.891	0.188	✓	-	-
Planar GS (ours)	**4.168**	**0.943**	**0.209**	✓	-	-

Replica. Both supervised methods of NeuralRecon and PlanarRecon trained on ScanNet cannot generalize to the new dataset, with NeuralRecon failing to produce valid meshes and PlanarRecon generating poor planar reconstruction results. On the other hand, our proposed PGS works well on Replica, with higher 3D plane segmentation scores and a lower inference time when comparing to the existing SOTA optimization-based method of NMF.

4.2. Ablation Study

We analyze different aspects of our proposed design. The ablation experiments are performed on two ScanNet scenes. More specifically, we study the effectiveness of (1) utilizing SAM masks for learning plane descriptors, (2) 2D normal maps supervision, (3) local planar alignment (Section 3.4), (4) applying holistic separability by using the recurrent mean-shift layer (Section 3.5), and (5) Laplacian smoothing of geometric features including normal and plane descriptor smoothing (Section 3.4). Table 3 shows the 3D plane segmentation performance on different variants of the proposed PGS, where we deactivate one component at each time. We observe a significant drop in performance when SAM masks are not used. This is expected, as the PGS learns plane descriptors using SAM, which are later used for grouping Gaussian nodes in tree and parsing plane instances. The absence of plane descriptors leads to high ambiguity in parsing individual small surfaces close to larger planar regions. Moreover, dropping normal vectors results in less degradation, as the plane descriptors using SAM masks can mainly resolve grouping ambiguity between Gaussian nodes in the tree structure. Furthermore, we see that the local planar alignment shows a high impact on the performance. This is because by enforcing the Gaussians to locate on local tangent planes of surfaces, it improves the learning of correct geometric features, such as normal vectors and plane descriptors, through rendering. While dropping such alignment results in a 3D reconstruction which point cloud is scattered around the surfaces. The Laplacian smoothing includes local averaging over both the normal and descriptor features of Gaussian primitives in training time, which further improves the performance. It

Table 3. Ablation Study on two ScanNet scenes.

Experiment	VOI↓	RI ↑	SC↑
W/o SAM masks	3.914	0.873	0.349
W/o local planar alignment	3.655	0.901	0.374
W/o normal vectors	3.326	0.904	0.390
W/o holistic separability	3.240	0.905	0.401
W/o Laplacian smoothing	3.151	0.908	0.393
Full Planar GS	**3.024**	**0.919**	**0.415**

can be seen that encouraging holistic separability also helps, providing effective performance improvement on plane segmentation.

5. Conclusions

In this paper, we proposed Planar Gaussian Splatting (PGS), which leverages two fundamental concepts: a probabilistic, hierarchical Gaussian mixture approach and a foundational vision model. Our approach represents the scene using a set of 3D planes, each defined by merging local geometries represented through 3D Gaussian distributions. To address ambiguity during the merging process, we introduce additional parameters to the Gaussians, including the normal vector and a plane descriptor. Learning plane descriptors without access to 2D/3D plane annotations involves utilizing a vision foundation model, specifically SAM. We learn 3D plane descriptors by constructing and partitioning a region adjacency graph based on SAM segments. Additionally, we address the challenges posed by variable-length and non-corresponding mask proposals across images via a linear regression approach. Experiments demonstrate that the proposed PGS outperforms existing competitive approaches in 3D planar reconstruction.

Limitations: The proposed method has some limitations. Dark regions in the image usually suffers under-reconstruction due to sparse assignment of Gaussian points. This can affect computing both KNN and the statistics such as mean and covariance of the point distributions on the plane. Additionally, very large planes in the scene might be split into two or several pieces as the likelihood of learning a compact descriptor for a large area containing a huge number of Gaussians primitives decreases. The proposed recurrent mean-shift updates mitigate this issue to some extent but does not resolve it completely.

References

[1] Samir Agarwala, Linyi Jin, Chris Rockwell, and David F Fouhey. PlaneFormers: From sparse view planes to 3D reconstruction. In *Proceedings of the European Conference on Computer Vision*, 2022. 1

[2] Matan Atzmon, Niv Haim, Lior Yariv, Ofer Israelov, Haggai Maron, and Yaron Lipman. Controlling neural level sets. *Advances in Neural Information Processing Systems*, 2019. 1

[3] Yash Bhalgat, Iro Laina, João F Henriques, Andrew Zisserman, and Andrea Vedaldi. Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion. *arXiv preprint arXiv:2306.04633*, 2023. 2, 3

[4] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, et al. Segment anything in 3d with nerfs. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[5] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. MobileNeRF: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3D reconstructions of indoor scenes. In *Proc. of the IEEE conf. on computer vision and pattern recognition*, 2017. 6

[7] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2022. 1

[8] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3D scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 4

[9] Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Nerf-sos: Any-view self-supervised object segmentation on complex scenes. *arXiv preprint arXiv:2209.08776*, 2022. 2

[10] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1, 6

[11] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2022. 2

[12] Jun Gao, Wenzheng Chen, Tommy Xiang, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Learning deformable tetrahedral meshes for 3D reconstruction. *Advances In Neural Information Processing Systems*, 33:9936–9947, 2020. 1

[13] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3D scene reconstruction with the manhattan-world assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 6

[14] Christian Hennig. Methods for merging gaussian mixture components. *Advances in data analysis and classification*, 4:3–34, 2010. 4

[15] Linyi Jin, Shengyi Qian, Andrew Owens, and David F Fouhey. Planar surface reconstruction from sparse views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 1, 3, 5, 6, 8, 11

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 4

[18] Shu Kong and Charless C Fowlkes. Recurrent pixel embedding for instance grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9018–9028, 2018. 6

[19] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. PlaneRCNN: 3D plane detection and reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 6

[20] Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. PlaneNet: Piece-wise planar reconstruction from a single RGB image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[21] Jiachen Liu, Pan Ji, Nitin Bansal, Changjiang Cai, Qingan Yan, Xiaolei Huang, and Yi Xu. PlaneMVS: 3D plane reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[22] Yichen Liu, Benran Hu, Junkai Huang, Yu-Wing Tai, and Chi-Keung Tang. Instance neural radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 787–796, 2023. 2

[23] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, 21(4):163–169, 1987. 1

[24] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[25] Ashkan Mirzaei, Yash Kant, Jonathan Kelly, and Igor Gilitschenski. Laterf: Label and text driven object radiance fields. In *European Conference on Computer Vision*, pages 20–36. Springer, 2022. 2

[26] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3D models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[27] Zak Murez, Tarrence Van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *Proceedings of the European Conference on Computer Vision*, pages 414–431, 2020. 12

[28] Yiming Qian and Yasutaka Furukawa. Learning pairwise inter-plane relations for piecewise planar reconstruction. In *Proceedings of the European Conference on Computer Vision*, 2020. 2

[29] Prafull Sharma, Ayush Tewari, Yilun Du, Sergey Zakharov, Rares Andrei Ambrus, Adrien Gaidon, William T Freeman,

Frédo Durand, Joshua B Tenenbaum, and Vincent Sitzmann. Neural groundplans: Persistent neural scene representations from a single image. In *International Conference on Learning Representations*, 2023. 2

[30] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3D shape synthesis. *Advances in Neural Information Processing Systems*, 2021. 1

[31] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9043–9052, 2023. 2

[32] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 6

[33] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 6, 8

[34] Bin Tan, Nan Xue, Song Bai, Tianfu Wu, and Gui-Song Xia. PlaneTR: Structure-guided transformers for 3d plane recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1, 2, 6

[35] Vadim Tschernezki, Diane Larlus, and Andrea Vedaldi. Neuraldiff: Segmenting 3d objects that move in egocentric videos. In *2021 International Conference on 3D Vision (3DV)*, pages 910–919. IEEE, 2021. 2

[36] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2022. 1

[37] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260*, 2021. 2

[38] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1

[39] Christopher Xie, Keunhong Park, Ricardo Martin-Brualla, and Matthew Brown. Fig-nerf: Figure-ground neural radiance fields for 3d object category modelling. In *2021 International Conference on 3D Vision (3DV)*, pages 962–971. IEEE, 2021. 2

[40] Yiming Xie, Matheus Gadelha, Fengting Yang, Xiaowei Zhou, and Huaizu 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*, 2022. 1, 2, 6, 7, 8, 12

[41] Fengting Yang and Zihan Zhou. Recovering 3D planes from a single image via convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[42] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 1

[43] Hong-Xing Yu, Leonidas J Guibas, and Jiajun Wu. Unsupervised discovery of object radiance fields. *arXiv preprint arXiv:2107.07905*, 2021. 2

[44] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. MonoSDF: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 2022. 1

[45] Zehao Yu, Jia Zheng, Dongze Lian, Zihan Zhou, and Shenghua Gao. Single-image piece-wise planar 3d reconstruction via associative embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[46] Farhad G Zanjani, Hong Cai, Yinhao Zhu, Leyla Mirvakhabova, and Fatih Porikli. Neural mesh fusion: Unsupervised 3d planar surface understanding. *arXiv preprint arXiv:2402.16739*, 2024. 1, 2, 6, 8

[47] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2