

OpenOcc: Open Vocabulary 3D Scene Reconstruction via Occupancy Representation

Haochen Jiang^{1*}, Yueming Xu^{1*}, Yihan Zeng², Hang Xu², Wei Zhang², Jianfeng Feng¹, Li Zhang^{1†}

Abstract—3D reconstruction has been widely used in autonomous navigation fields of mobile robotics. However, the former research can only provide the basic geometry structure without the capability of open-world scene understanding, limiting advanced tasks like human interaction and visual navigation. Moreover, traditional 3D scene understanding approaches rely on expensive labeled 3D datasets to train a model for a single task with supervision. Thus, geometric reconstruction with zero-shot scene understanding i.e. Open vocabulary 3D Understanding and Reconstruction, is crucial for the future development of mobile robots. In this paper, we propose OpenOcc, a novel framework unifying the 3D scene reconstruction and open vocabulary understanding with neural radiance fields. We model the geometric structure of the scene with occupancy representation and distill the pre-trained open vocabulary model into a 3D language field via volume rendering for zero-shot inference. Furthermore, a novel semantic-aware confidence propagation (SCP) method has been proposed to relieve the issue of language field representation degeneracy caused by inconsistent measurements in distilled features. Experimental results show that our approach achieves competitive performance in 3D scene understanding tasks, especially for small and long-tail objects.

I. INTRODUCTION

Reconstructing 3D scenes with high-level semantics from digital images and videos is a crucial area of research in 3D computer vision and robotics, which has been widely used in various fields such as visual navigation [1], [2], and robotic grasping [3], [4]. However, the intrinsic connections and underlying homologies between scene reconstruction and understanding have been greatly neglected. These two tasks are often treated as separate studies, where the bottleneck lies.

Several commonly used 3D representations include 3D point clouds [5], [6], voxel grids [7], [8], and polygonal meshes [9]. Directly learning semantics on these 3D representations requires a substantial amount of well-annotated 3D ground truths for training. It is more laborious and expensive to obtain the 3D ground truth label. Moreover, the majority of existing methods [10]–[13] encounter challenges in achieving model transferability owing to domain gaps. Recently, neural radiance fields (NeRFs) [14] have demonstrated the power of implicit representation in intricate real-world 3D scenes, which can capture intricate geometric details using solely RGB images. This novel scene representation is more compact, continuous, efficient, and able to be optimized with

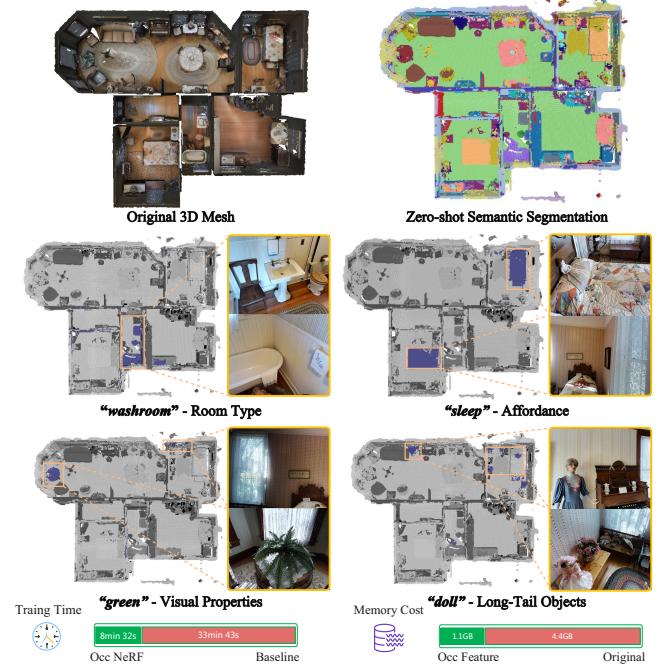


Fig. 1. Open-vocabulary 3D scene understanding and reconstruction. We propose OpenOcc, a zero-shot method for 3D scene structure perception. The examples above show zero-shot 3D scene semantic segmentation results with the occupancy feature grid. The blue color denotes the matching results of a user-specified query string to demonstrate the flexibility of the language-based features grid.

differentiable rendering. Many researchers utilize this 3D representation to conduct the Neural Semantic fields [10], [15]. However, these existing works do not incorporate geometric (depth) and multi-view constraints during the training process, leading to the emergence of shape-radiance ambiguity like LeRF [16]. Moreover, these methods can not effectively reconstruct geometric structures in large-scale scenes, like Matport3D [17], and lack the open vocabulary capabilities to understand environments. Most reconstruction works utilize this 3D representation [18], [19] to achieve high-fidelity and multi-view consistent reconstructions with reduced memory usage. However, the reconstructed geometric structure lacks the incorporation of high-level semantics. Hence, we ask a natural question:

Can we customize a better way to effectively integrate the benefits of scene reconstruction and open vocabulary 3D scene understanding for generating a consistent semantic map?

However, it is not feasible to directly fuse these two types of work for the robotic structure perception, owing to two

*Haochen Jiang and Yueming Xu contributed equally to this work.

¹Haochen Jiang, Yueming Xu, Jianfeng Feng, and Li Zhang are with Fudan University. ([†]Li Zhang is the corresponding author: lizhangfd@fudan.edu.cn)

²Yihan Zeng, Hang Xu and Wei Zhang are with Huawei Noah's Ark Lab.

facts: 1) Current reconstruction methods [18], [20], [21] only perform better on 3D object surface reconstruction, which can not be extended to complex or large-scale indoor scenes. 2) Owing to the potential noise of language embedding, the open vocabulary segmentation predictions are inconsistent and ambiguous. Directly distilling these inconsistent features may lead to the degeneracy of the representation capabilities for language fields, which causes floaters in volume rendering.

To tackle this question, we propose *OpenOcc*, a novel framework unifying the scene reconstruction and open vocabulary understanding utilizing neural radiance field. To better learn the implicit representation, we construct the language, geometric, and appearance fields with separate multi-resolution feature grid. Instead of predicting a sign distance function (SDF), we utilize the occupancy representation surface as the decision boundary of a binary occupancy classifier. This leads to a significant reduction in memory requirements and improves the inference efficiency compared to the dense point-based and mesh-based representation. Moreover, the small objects and thin structures can be better reconstructed. To achieve open vocabulary scene understanding, we distill the 2D language embeddings into a 3D language field inspired by LeRF [16]. To address the issue of shape-radiance ambiguity, we have developed geometric constraints that incorporate depth information. Owing to the potential noise in the language embedding, the predictions from open vocabulary segmentation exhibit inconsistencies and ambiguities. Thus, we propose an SCP method to dynamically reallocate the weights of the corresponding language features in the language field updating, which can significantly improve the accuracy of 2D segmentation results.

In summary, our **contributions** are summarized as follows:

- We design a novel open vocabulary reconstruction system, unifying the 3D scene reconstruction and understanding with neural radiance fields, which can simultaneously achieve geometric reconstruction and zero-shot perception from a set of posed RGB-D images.
- We introduce a novel approach of semantic-aware confidence propagation (SCP) within the process of updating semantic radiance field features. This method addresses the challenge of inconsistent open vocabulary segmentation across diverse viewpoints.
- We demonstrate that the proposed methods can be used for 3D semantic segmentation with better performance than state-of-the-art zero-shot methods, particularly for small and long-tail objects.

II. RELATED WORK

a) Neural Implicit 3D Scene Reconstruction: Neural implicit scene representations, also known as neural fields [14], have gained considerable attention in the field of 3D geometry reconstruction owing to their high level of expressivity and minimal memory requirements. NeuS [22] introduces a volume rendering method to train an unbiased neural Signed

Distance Function (SDF) representation. VolSDF [23] proposes a parameterization method to transfer the volume density to SDF. UniSurf [20] unifies neural volume and surface rendering, enabling both within the same model. These works only perform better on 3D object reconstruction, which can not expand to complex indoor scenes. Thus, there occurs some works [24], [25], exploiting geometric priors to improve multi-view 3D scene reconstruction quality, efficiency, and scalability. Many works follow this framework and improve it in many aspects, such as model acceleration [26], [27], and reconstruction precision [28]–[30]. Considering the critical factors of efficiency, reconstruction quality, and integrity of 3D scene depiction for robotic navigation, we exploit a more practical occupancy representation instead of SDF to reconstruct the indoor scenes inspired by [18], [19].

b) Open-Vocabulary 2D Scene Understanding: Visual Language Models (VLMs) [31], [32] have recently achieved the capability to establish robust mappings between images and textual content, leading to a remarkable level of performance in zero-shot 2D scene understanding tasks. However, VLMs only provide image-level embeddings, which can not apply to the dense prediction task requiring pixel-level information such as segmentation. Thus, many works [33]–[37] attempt to align the per-pixel image feature with the large language embeddings. In this way, users can give an arbitrary text label and query the interest objects on-the-fly. Compared with the close-set 2D scene understanding methods [38], [39], we exploit a pre-trained open-vocabulary 2D models [33], [34] and distill the 2D feature into 3D to finish a series of 3D understanding tasks, without any ground truth label in the whole training process. Leveraging the robust comprehension power of an open vocabulary, our approach demonstrates proficiency in recognizing long-tail objects and those objects that lack label annotations within the dataset. This effectively facilitates the open-ended perception potential crucial for autonomous robotic systems.

c) Grounding Language into 3D Representations: Neural implicit feature fields have emerged as a method to translate 2D features into a grounded 3D representation. CLIP-Fields [40] fuse CLIP [31] embeddings of crops into point clouds using a contrastively supervised field. Open-Scene [41] and VL-Maps [2] build a 3D volume of language features aligned with the pre-trained language segmentation encoders [34], [35]. These methods are significantly constrained as they do not encode the intrinsic scene geometry, relying instead on an external point cloud for query operations. LERF [16] and VL-Fields [42] advocate the construction of a language field through volume rendering and the incorporation of the 2D language feature [31], [34] into the 3D implicit field, which is beneficial to obtain the dense and continuous scene perception results. However, owing to the potentially noisy nature of language embeddings, the predictions from open vocabulary segmentation exhibit inconsistencies and ambiguity. In this paper, we propose a semantic-aware confidence propagation method to dynamically reallocate the weights of the corresponding

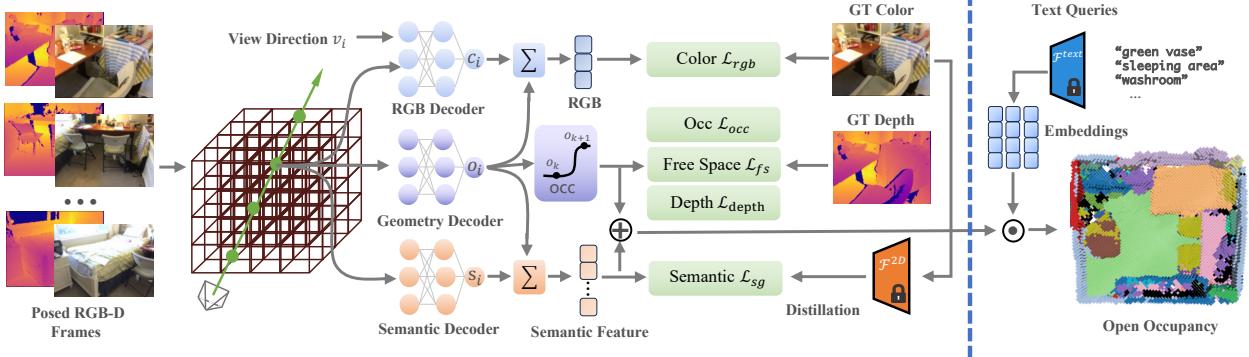


Fig. 2. **The overall framework of the proposed method.** *Left:* Given a series of posed RGB-D frames, we construct the RGB, geometry, and semantic decoders via separate multi-resolution feature grid with the geometric loss \mathcal{L}_{occ} , \mathcal{L}_{fs} , \mathcal{L}_{depth} and color loss \mathcal{L}_{color} . To learn the language knowledge, we distill the dense feature \mathcal{F}_{2d} via volume rendering with a distillation loss \mathcal{L}_{sg} . *Right:* During inference, we can compute the similarity score between user's text embeddings and generate an occupancy feature map to perform the open-vocabulary 3D understanding task.

language features in the language field updating. The mistake observation of language embeddings can be efficiently filtered.

III. METHODOLOGY

The overall architecture of the proposed methods is illustrated in Fig. 2. The entire framework can be divided into three parts. Firstly, an explicit occupancy neural radiance field is trained to reconstruct the geometric structure of the environments. Secondly, we introduce a semantic language field (SLF) utilizing the open-vocabulary semantic segmentation methods such as OpenSeg [35] to distill the 2D features into 3D space, which improves the robotics ability of 3D scene understanding. Finally, during inference, we can achieve an open vocabulary segmentation task by computing the similarity scores between an arbitrary text query and a generated semantic feature map with occupancy representation.

A. Occupancy Field Generation

a) Multi-resolution Feature Grid: To better reconstruct details of the 3D scene geometry structure, we utilize the multi-resolution feature grid via occupancy representation instead of a single scale inspired by Go-Surf [26]. We encode the scene geometry into a multi-level feature grid $\mathbb{V}_\theta = \{V^l\}$, where $l \in \{0, 1, 2, \dots, L\}$ encode the geometry details from coarse to fine. The coarse-level features can fastly initialize scene geometry presentation to fill the hole. Furthermore, the geometry residual of the high-frequency details will be refined by the fine-level features grid.

For a random sample point $\mathbf{p} \in \mathbb{R}^3$ in a scene, we query the representation feature by tri-linear interpolation operation at each level and concat the multi-level features as the input of the decoder to obtain the occupancy probability $o(p)$ via the MLP $f_\alpha(\cdot)$:

$$\mathcal{V}_\theta(\mathbf{p}) = [V^0(\mathbf{p}), V^1(\mathbf{p}), V^2(\mathbf{p}), \dots, V^L(\mathbf{p})], \quad (1)$$

$$o(\mathbf{p}) = f_\alpha(\mathcal{V}_\theta(\mathbf{p})). \quad (2)$$

To reduce the ambiguity between geometry and appearance in the learning process, we use a separate feature grid

$\mathcal{C}_\phi, \mathcal{S}_\omega$ to module the color and semantic fields, respectively. Similarly, we use different decoders $g_\beta(\cdot), h_\gamma(\cdot)$ to infer the color and semantic features.

$$c(\mathbf{p}, \mathbf{d}) = g_\beta(\mathcal{C}_\phi(\mathbf{p}), \mathbf{d}), \quad (3)$$

$$s(\mathbf{p}) = h_\gamma(\mathcal{S}_\omega(\mathbf{p})), \quad (4)$$

where \mathbf{d} is the view direction of the sample ray. Here θ, ϕ, ω and α, β, γ represent the optimization parameters about grid features and MLP weights, respectively.

b) Volume Rendering with Occupancy Field: To obtain the explicit feature grid indirectly, we adopt an occupancy field instead of a density field to represent the scene's geometry structure inspired by [18] and [19]. Instead of predicting a discrete voxel representation, we predict the continuous occupancy function to generate an occupancy field which is similar to SDF representation. Given a ray \mathbf{r} from the virtual camera's center connected with a random pixel $[u, v]$, we sample a set of N 3D points $\{\mathbf{x}_i \in \mathbb{R}^3\}_N$ along the viewing direction $\mathbf{d} \in \mathbb{R}^3$. Following the Eq. 3, we can obtain the predicted RGB color value $c_i(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^3$ at each position \mathbf{x} . For every location \mathbf{x} , we predict an occupancy probability between 0 and 1 via a neural network as Eq. 5.

$$o(\mathbf{p}) : \mathbb{R}^3 \longrightarrow \{0, 1\} \quad (5)$$

where $o(\mathbf{p}) = 0$ and $o(\mathbf{p}) = 1$ denotes the points located in free space and occupied space, respectively. Note that we are interested in the decision boundary, which implicitly represents the scene's surface with a continuous occupancy field. According to the physical implications of accumulated transmittance $T_i = \exp\left(-\sum_{j < i} \sigma(\mathbf{x}_j) \delta_j\right)$, the rendering equation of color and depth can be rewritten as :

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^N o(\mathbf{x}_i) \prod_{j < i} (1 - o(\mathbf{x}_j)) c(\mathbf{x}_i, \mathbf{d}), \quad (6)$$

$$\hat{D}(\mathbf{r}) = \sum_{i=1}^N o(\mathbf{x}_i) \prod_{j < i} (1 - o(\mathbf{x}_j)) z_i,$$

where the terms $\prod_{j < i} (1 - o(\mathbf{x}_j))$ denotes the visible probability which is 1 if there is no occupied sample \mathbf{x}_j with

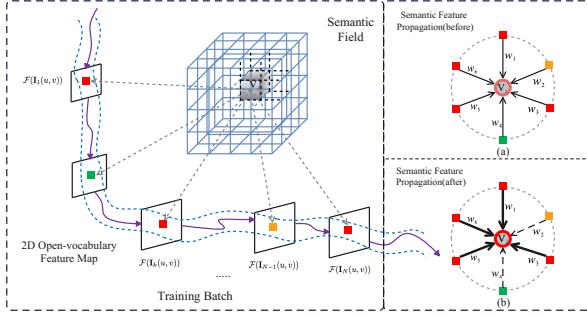


Fig. 3. **Different semantic feature update strategies.** Owing to the potentially noisy language embedding, the open vocabulary segmentation results in different views are inconsistent in the same training batch that can show the left image. (a) Semantic field updates use the same weight. (b) The proposed Semantic-aware Confidence Propagation (SCP). Dashed lines mean smaller weight. The boundary color of the center point means the fusing feature is dominated by which consistent semantic class.

$j < i$ before sample \mathbf{x}_i . Moreover, $o(\mathbf{x}_i) \prod_{j < i} (1 - o(\mathbf{x}_j))$ represents the ray first hit the surface at \mathbf{x}_i which the occupied value is 1, and other samples is 0. Thus, $\hat{C}(\mathbf{r})$ and $\hat{D}(\mathbf{r})$ takes the color $c(\mathbf{x}_i, \mathbf{d})$, depth $z(\mathbf{x}_i)$ of the first occupied sample \mathbf{x}_i along the ray, respectively.

B. Multi-view Consistent Language Fields

a) Image Feature Extraction: Given an RGB image \mathbf{I}_i with a resolution of $H \times W$, we can obtain the per-pixel embedding from a pre-trained visual language segmentation model $\mathcal{F}(\cdot)$ as :

$$\mathcal{F}(k_i) : \mathbb{R}^3 \longrightarrow \mathbb{R}^D, k_i \in I[u, v] \quad (7)$$

where $\mathcal{F}(\mathbf{I}_i) \in \mathbb{R}^{H \times W \times D}$, D is the language feature dimension, and i is the index sample from a image sequences.

b) Features Distillation: To obtain the queryable scene representation, we distill the 2D visual-language feature into 3D explicit semantic fields utilizing volume rendering inspired by [10], [16], [41]. Given a position \mathbf{x}_i in space, we can query the volume feature and decode the semantic feature as Eq. 4. Similarly, the expected semantic feature $\hat{S}(\mathbf{r})$ along each pixel/ray can be defined as :

$$\hat{S}(\mathbf{r}) = \sum_{i=1}^N o(\mathbf{x}_i) \prod_{j < i} (1 - o(\mathbf{x}_j)) s(\mathbf{x}_i), \quad (8)$$

where the dimension of $\hat{S}(\mathbf{r})$ is D , which is equal and corresponding to $\mathcal{F}(k_i)$. To keep consistent with the CLIP feature, we normalize each raw output to the unit sphere as $S(\mathbf{r}) = \hat{S}(\mathbf{r}) / \|\hat{S}(\mathbf{r})\|$. Since the open-vocabulary segmentation embedding from [34], [35] is co-embedded with CLIP feature, the distilled 3D semantic language field is also located in the same space as CLIP naturally.

c) Semantic-aware Confidence Propagation (SCP): To obtain a more consistent and accurate semantic representation within the language field, we propose a probabilistic fusion approach, denoted as SCP, to integrate the multi-view semantic features as shown in Fig. 3, which can dynamically adjust the weights of the features during the language field

updating. During a discrete update interval at time t , each feature grid is associated with N measurements in a batch, represented as $z_t(N)$. Denote the k -th label probability distribution of a feature grid at the current optimization step as $s_t(k)$, we consider the history sampling measurements till now as $z_{1:t}$ and then update the 3D semantic language fields based on the Bayes' rule:

$$P(s_t | z_{1:t}) = \frac{P(s_t | z_t) P(z_t)}{P(s_t)} \frac{P(s_{t-1} | z_{1:t-1})}{P(z_t | z_{1:t-1})}. \quad (9)$$

The opposite event \tilde{s}_t of posterior distribution can be obtained by:

$$P(\tilde{s}_t | z_{1:t}) = \frac{(1 - P(s_t | z_t)) P(z_t) (1 - P(s_{t-1} | z_{1:t-1}))}{(1 - P(s_t)) P(z_t | z_{1:t-1})}. \quad (10)$$

Dividing the above two equations Eq. 9 and Eq. 10. We utilize the log-odds algorithm from Thrun [43] to quantify the update weights for semantic labels:

$$l_t(s) = l(s_t | z_t) + l(s_{t-1} | z_{1:t-1}) - l_0, \quad (11)$$

where $l_t(s) = \log \frac{P(s_t | z_{1:t})}{1 - P(s_t | z_{1:t})}$ denotes the probability of the semantic class at time t . All semantic classes are initially assigned equal probabilities in each feature grid, which indicates $l_0 = 0$. During the practical computation process, at each time step t , the state s_t typically has k semantic classes in different datasets. The observation likelihood for the k -th class, $P(s_t(k) | z_t)$, is calculated as the ratio N_k/N . Furthermore, the normalized confidence weight w_k for each semantic label is derived by corresponding $l_t(s_k)$ as

$$w(s_k) = \frac{l_t(s_k)}{\sum_{j=1}^K l_t(s_j)} \quad (12)$$

C. Training Loss

a) Color and Depth Rendering Loss: To obtain the essential geometry presentation of neural implicit field, we render depth and color in independent view as Eq. 6 comparing with the proposed ground truth (GT) map:

$$\begin{aligned} \mathcal{L}_{rgb}(\mathbf{r}) &= \frac{1}{M} \sum_{i=1}^M \|\hat{C}(\mathbf{r}_i) - C(\mathbf{r}_i)\|_2^2. \\ \mathcal{L}_{depth}(\mathbf{r}) &= \frac{1}{N_d} \sum_{\mathbf{r} \in N_d} \|\hat{D}(\mathbf{r}) - D(\mathbf{r})\|_2^2. \end{aligned} \quad (13)$$

Where $C(\mathbf{r})$ and $D(\mathbf{r})$ denote the ground truth color and depth corresponding with the given pose, respectively. M represents the number of sampled pixels in the current image. Note that only rays with valid depth value N_d are considered in \mathcal{L}_{depth} .

b) Approximate Occupancy Supervision: Directly optimizing the occupancy field from RGB-D observation is not sufficient to reconstruct the high-frequency details and may cause the shape-radiance ambiguity issue. Inspired by SDF constraints in [24], [26], we introduce occupancy supervision with the binary cross-entropy (BCE) loss between the predicted $\hat{o}_{\mathbf{x}_i}$ and the true occupancy values $o_{\mathbf{x}_i}$ as:

$$\mathcal{L}(\hat{o}_{\mathbf{x}_i}, o_{\mathbf{x}_i}) = -[o_{\mathbf{x}_i} \cdot \log(\hat{o}_{\mathbf{x}_i}) + (1 - o_{\mathbf{x}_i}) \cdot \log(1 - \hat{o}_{\mathbf{x}_i})] \quad (14)$$

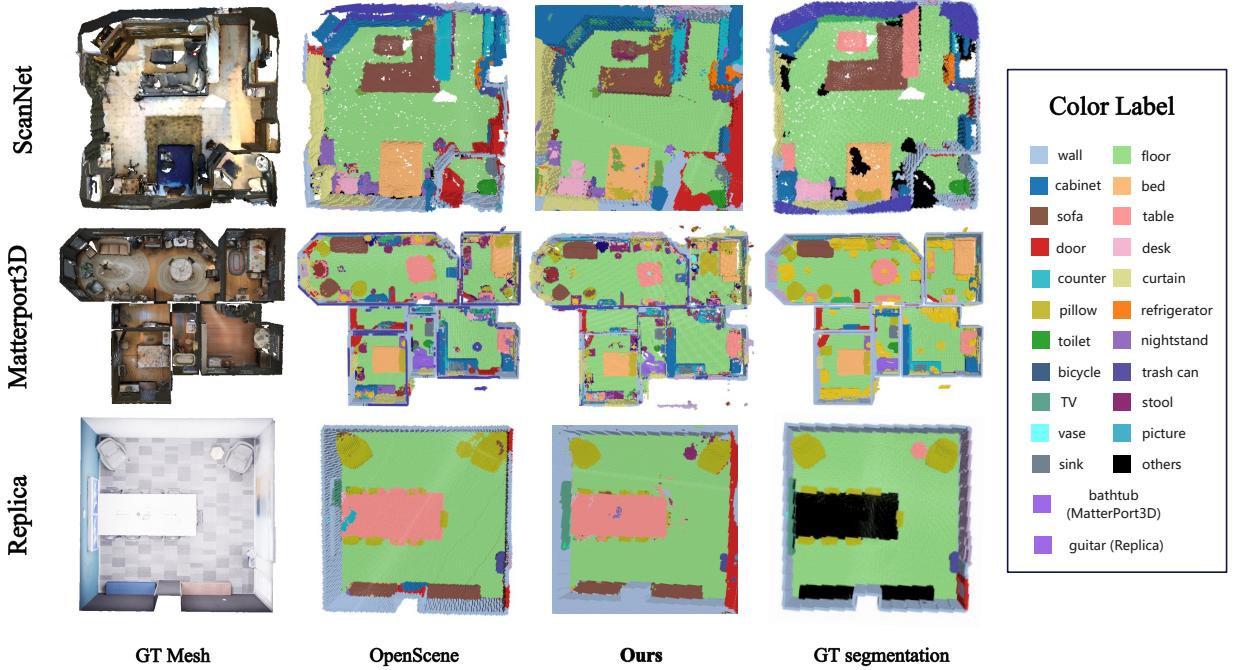


Fig. 4. **Qualitative comparisons.** Images of 3D semantic segmentation results on three public indoor benchmarks.

For the occupancy representation, the surface will exist the thickness owing to the grid resolution. We suppose that the point of occupancy probability, which is located near the surface as $|D(\mathbf{x}_i) - d| \leq t$, is 1. The bound will be controlled by a truncation threshold t (we set $t = 5$ cm).

$$\mathcal{L}_{occ}(\hat{o}_{\mathbf{x}_i}, o_{\mathbf{x}_i}) = \frac{1}{M} \sum_{i=1}^M \frac{1}{|\mathcal{R}_{tr}|} \sum_{i \in \mathcal{R}_{tr}} \mathcal{L}(\hat{o}_{\mathbf{x}_i}, o_{\mathbf{x}_i}) \Big|_{o_{\mathbf{x}_i}=1} \quad (15)$$

For the depth value out of the truncation, we apply the \mathcal{L}_{fs} loss to encourage the occupancy probability of free space near to 0.

$$\mathcal{L}_{fs}(\hat{o}_{\mathbf{x}_i}, o_{\mathbf{x}_i}) = \frac{1}{M} \sum_{i=1}^M \frac{1}{|\mathcal{R}_{fs}|} \sum_{i \in \mathcal{R}_{fs}} \mathcal{L}(\hat{o}_{\mathbf{x}_i}, o_{\mathbf{x}_i}) \Big|_{o_{\mathbf{x}_i}=0} \quad (16)$$

c) 2D-3D Semantic Feature Aligned Loss: To enforce the semantic field aligned with the 2D open vocabulary feature, we use a cosine similarity loss as:

$$\mathcal{L}_{sg}(\mathbf{r}) = \frac{1}{M} \sum_{i=1}^M \rho(w_i \cdot \|1 - \cos(\mathcal{F}(\mathbf{r}_i), \hat{S}(\mathbf{r}_i))\|_2^2) \quad (17)$$

where w_i denotes the normalized confidence weight of the corresponding semantic feature, which has been mentioned in Sec. III-B. In addition, $\rho(\cdot)$ is a robust kernel function to restrain the effect of the semantic outlier features. The δ is a hyper-parameter of Huber loss (we set $\delta = 1$). Owing to this robust kernel, the label lived in edge can be more exact. We show the comparison results in the ablation study in Sec. IV-E.

The overall loss function is finally formulated as the following minimization problem:

$$\begin{aligned} \min_{\mathcal{P}} & \lambda_1 \mathcal{L}_{rgb}(\mathbf{r}) + \lambda_2 \mathcal{L}_{depth}(\mathbf{r}) + \lambda_3 \mathcal{L}_{occ}(\mathbf{r}) + \lambda_4 \mathcal{L}_{fs}(\mathbf{r}) \\ & + \lambda_5 \mathcal{L}_{sg}(\mathbf{r}), \end{aligned} \quad (18)$$

where $\mathcal{P} = \{\theta, \phi, \omega, \alpha, \beta, \gamma, \xi_i\}$ is the list of parameters being optimized, including fields feature, decoders, and camera pose. The weights of each loss item are set $\lambda_1 = 10.0$, $\lambda_2 = 1.0$, $\lambda_3 = 10.0$, $\lambda_4 = 1.0$, $\lambda_5 = 2.0$.

D. Zero-shot Inference

To achieve open vocabulary 3D understanding for robotics visual navigation tasks, we combine the semantic field and occupancy field $\mathcal{Q}(\cdot)$ into a new explicit feature map under a fixed resolution.

$$\mathcal{Q}(\mathbf{x}_i) = \mathcal{V}_\theta(\mathbf{x}_i) \otimes \mathcal{S}_\omega(\mathbf{x}_i), \quad (19)$$

where \otimes denotes an operation that selects the semantic feature in which the occupancy probability is 1 at position \mathbf{x}_i . Given a set of language tokens $\mathbf{t}_i \in \mathcal{T}$, we transfer these words into prompts and utilize a text encoder \mathcal{E} to obtain text features. The final segmentation results for each 3D point \mathbf{x}_i can be computed point-wise by:

$$s(\mathbf{x}_i) = \text{argmax}_i \{ \cos(h_\gamma(\mathcal{Q}(\mathbf{x}_i)), \mathcal{E}(\mathbf{t}_i)) \} \quad (20)$$

IV. EXPERIMENT

A. Experimental Setup

a) Datasets: We train and evaluate our model mainly on three popular public benchmarks: Replica [44], ScanNet [45], and Matterport3D [17]. These three datasets encompass the major situations of indoor scenes: the first two represent regular indoor scenes, and the last represents complex and large-scale indoor scenes, which all provide a series of posed RGB-D images. To more accurately and authentically showcase the segmentation capabilities of these methods on previously unaddressed classes, we also evaluate our method on the Scannet-200 dataset [46].

TABLE I

CLASSWISE SEMANTIC SEGMENTATION RESULTS ON THE SCANNET-200 [46] VALIDATION SET. WE SHOW THE QUANTITATIVE COMPARISON BETWEEN OUR METHOD AND THE MOST RECENT ZERO-SHOT 3D SEGMENTATION APPROACH: OPENSCENE [41].

	mIoU				mAcc			
	Clock	Light	Fan	Book	Clock	Light	Fan	Book
OpenScene	37.6	1.7	6.4	1.0	37.7	53.8	25.9	1.0
Ours	52.9	8.0	29.9	19.6	66.1	75.9	32.8	25.9

TABLE II

COMPARISONS ON INDOOR BENCHMARKS. WE COMPARE OUR METHOD WITH THE ZERO-SHOT BASELINES FOR 3D SEMANTIC SEGMENTATION ON INDOOR BENCHMARKS.

	Replica [44]		ScanNet-200 [46]		Matterport [17]	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
MSeg Voting [49]	-	-	-	-	33.4	39.0
OpenScene [41]	46.9	69.0	13.3	25.7	42.5	60.4
Ours	50.5	65.4	17.5	26.8	45.1	63.3

b) Metrics: The evaluation of the reconstructed 3D structure involves assessing six standard metrics as defined in [47]: Accuracy (Acc), Completeness (Comp), Chamfer- \mathcal{L}_1 (C-L_1), Precision (Prec), Recall, and F-score. The F-score is computed using a threshold of 5cm. For scene understanding, we compute the two most common evaluation metrics, mean intersection-over-union (mIoU) and mean accuracy (mAcc), which are used to evaluate the segmentation performance and accuracy of models. There is no unified and accurate evaluation metric for open vocabulary scene perception, so we mainly perform qualitative comparisons and discussions to demonstrate the zero-shot capability of our proposed method.

c) Implementation Details: We run our OpenOcc on a Tesla A6000 GPU for about 30 to 40 minutes(depending on the scene size) to finish the reconstruction and perception, which takes roughly 30GB of memory in total. For all our experiments, we sample 6144 rays per patch and sample 132 points along each ray. We further train our model for 10k iterations with a learning rate of 1×10^{-2} and 1×10^{-3} for MLP decoders and feature grids. Due to the universality of our approach, unconstrained by specific semantic segmentation methods, we choose pre-trained OpenSeg [48] for 2D feature distillation in our experiment part.

B. Experimental Results

a) Results on 3D semantic segmentation: We evaluate our OpenOcc on all classes of the Replica, Scannet-200 validation set and Matterport3D test set. In Tab. I and Tab. II, we compare our methods with the most closely related work on zero-shot methods. Due to the inherent shape-radiance ambiguity problem associated with the LeRF approach, it can not acquire accurate geometric structures on the validation dataset. Consequently, a comparison with it has not been conducted. Our model integrates 3D scene reconstruction with open vocabulary segmentation, exhibit-

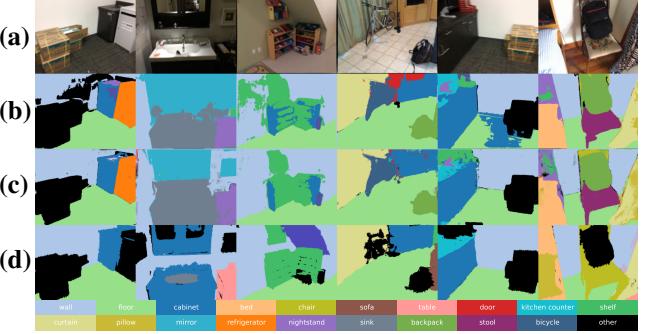


Fig. 5. 2D segmentation results on ScanNet. We visualized some 2D segmentation examples from the ScanNet validation set. (a) depicts the Input Image, while (b) showcases the OpenSeg result, (c) illustrates our method, and (d) represents GT Segmentation. Black pixels in the ground truth segmentation correspond to classes not included in the ScanNet-20 evaluation classes.

ing superior performance relative to OpenScene, particularly in the segmentation of small objects or long-tail objects, as evidenced in Tab. I. Since reconstructing a large number of scenes will consume too much resource and time, we establish a new validation subset by randomly extracting 10% of data(about 30 scenes) from the Scannet validation set, and a similar operation is performed on the Matterport3D test set. For fairness, we also evaluate OpenScene under the same subset split.

As shown in Tab. II, our model outperforms the zero-shot baselines (OpenScene) in terms of mIoU and mAcc on Matterport3D and ScanNet-200 and achieves a higher mIoU score in Replica. Owing to the motion blur and invalid depth value of collected images on ScanNet, it is a challenge for most reconstruction methods and will significantly reduce the reconstruction quality. In contrast, OpenScene extracts the point cloud from the GT mesh, which undergoes post-processing operations. This leads to variances in the quality of the reconstructed geometry. However, we demonstrate that our method still achieves superior results in classes previously unaddressed, as evidenced by our findings on the ScanNet-200 dataset presented in Tab. I. Owing to the SCP method relieving the issue of language field representation degeneracy, our approach exhibits enhanced segmentation capability for small objects, including vases, pillows, and stools, as illustrated in Fig. 4. This finding aligns with the conclusions presented in Tab. I.

b) Results on 2D semantic segmentation: Compared to other 3D segmentation methods, our method can also obtain any view of 2D segmented results owing to the view synthesis capabilities of NeRFs. Visual comparisons of 2D semantic segmentations on ScanNet are shown in Fig. 5. Despite errors or inconsistencies in the 2D segmentation from OpenSeg, our method can still correctly segment and complete the targets in the scene through volume rendering. For example, the OpenOcc successfully segmented the wall in the first column of Fig. 5 and completely identified the floor in the fifth column of Fig. 5. This demonstrates that our method can better fuse the language embeddings and obtain

TABLE III

QUANTITATIVE EVALUATION OF RECONSTRUCTION MAP. WE EVALUATE THE 3D RECONSTRUCTION PERFORMANCE OF OUR PROPOSED METHOD AGAINST PREVIOUS WORKS ON SCANNET DATASET WITH SIX REPRESENT METRICS. “*” DENOTES THE RESULTS GENERATED FROM THE SOURCE CODE WITH THE DEFAULT PARAMETERS.

Method	Acc↓	Comp↓	C-L ₁ ↓	Prec↑	Recall↑	F-score↑
COLMAP [50]	0.047	0.235	0.141	71.1	44.1	53.7
UNISURF [20]	0.554	0.164	0.359	21.2	36.2	26.7
VolSDF [23]	0.414	0.120	0.267	32.1	39.4	34.6
NeuS [22]	0.179	0.208	0.194	31.3	27.5	29.1
Manhattan-SDF [29]	0.072	0.068	0.070	62.1	56.8	60.2
NeuRIS [51]	0.050	0.049	0.050	71.7	66.9	69.2
MonoSDF [24]	0.035	0.048	0.042	79.9	68.1	73.3
OccSDF [30]	0.039	0.041	0.040	80.0	76.0	77.9
Go-Surf* [26]	0.102	0.024	0.063	79.0	91.2	84.2
Ours	0.044	0.029	0.037	81.2	90.0	85.3

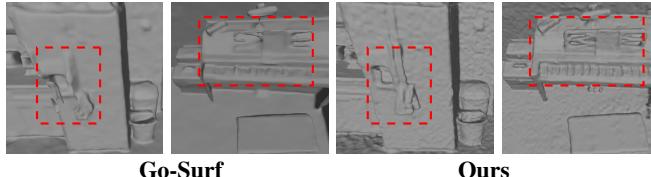


Fig. 6. **Qualitative results of detailed reconstructed architectures.** Compared with our baseline Go-Surf [26], our approach demonstrates superior accuracy in reconstructing shapes and contours of objects on Scannet scene 0 (left) and scene 50 (right).

more consistent, complete, and accurate 2D segmentation results.

c) Reconstruction Quality: To better demonstrate the performance of our proposed occupancy representation in scene reconstruction, we evaluate the reconstructed 3d structure from both qualitative and quantitative perspectives on ScanNet. We use the test split from [29] and uniformly sample roughly 120 frames to train our network. In our framework, as we generate an occupancy grid, we extract meshes based on the occupancy probability using marching cubes with a resolution of 1cm and a threshold of 0.5. Regions that are not observed in any camera views are culled before evaluation. In Tab. III, we conducted a comparison of our proposed approach against existing implicit reconstruction methods, including Manhattan-SDF [29], MonoSDF [24], Go-Surf [26] and OccSDF [30]. Our approach achieves comparable performance in 3D scene reconstruction. Compared with our baseline method Go-Surf [26], our approach demonstrates high-quality reconstructions of shapes and contours for objects, as illustrated in Fig. 6.

C. Application of robotic visual navigation

To further illustrate the potential applicability and efficiency of OpenOcc within the area of mobile robotics, we have conducted a simulation experiment for visual navigation tasks. This experiment is designed to highlight the advantages of the proposed method for practical robotic navigation, as evidenced in Fig. 7. Upon receiving navigation instructions, leveraging ChatGPT finishes task decomposition, which aids in acquiring detailed information about the origin (kitchen sink) and destination objects (washroom

TABLE IV

ABLATION STUDY. COMPARISON OF SEMANTIC SEGMENTATION PERFORMANCE OF DIFFERENT PROPOSED COMPONENTS BY OUR METHOD.

	mIoU	mAcc
w/o Huber	44.6	62.6
w/o SCP	43.7	62.8
w/o BCE	44.0	59.6
OpenOcc	45.1	63.3

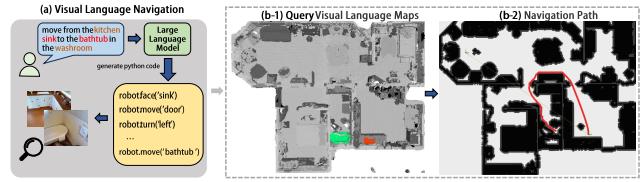


Fig. 7. **The simulation experiments in robotic visual navigation.**

bathtub). By querying our visual language map, we can obtain precise location information, as illustrated in part b-1 of Fig. 7. Subsequently, a navigation route is devised utilizing the path planning module within ROS2, as depicted in part b-2 of Fig. 7. The open-vocabulary occupancy map we have reconstructed facilitates mobile robots in more effectively completing visual navigation tasks.

D. Memory and Time

We exhibit the merits of our occupancy-based reconstruction and representation method in terms of low storage requirements and high efficiency, shown in Fig. 1. Compared with the SDF-based NeRF method, our occupancy-based NeRF variant reduces training time by a significant 74.7%, from 33min down to just 8min. Moreover, implicit fields typically require 4.4GB of memory, whereas our explicit language feature occupancy grid takes only 1.1GB of memory, a reduction of 75%.

E. Ablation Study

To verify the effectiveness of our network design, including the Huber loss, Binary Cross Entropy (BCE) loss, and the Semantic-aware Confidence Propagation (SCP) module, we test the performance of our method under different settings. The results presented in Tab. IV demonstrate that all the proposed strategies are effective on matterport3D segmentation. This suggests that adopting the BCE loss can promote better occupancy-based reconstruction. At the same time, leveraging SCP and the robust kernel can effectively eliminate training instability caused by inconsistent 2D open segmentation.

V. CONCLUSION

We present OpenOcc, a novel framework that integrates 3D scene understanding and reconstruction using Neural Radiance Fields, which combines the advantage of the computationally efficient occupancy representation and the open vocabulary scene understanding capability. Experimental results demonstrate our method can reconstruct high-quality geometric structures and exhibit competitive performance

on three public benchmarks compared with current state-of-the-art scene understanding methods. The supplementary external simulation experiment further underscores the prospective applicability of our proposed method for robotic visual navigation.

REFERENCES

- [1] C. Galindo, J.-A. Fernández-Madrigal, J. González, and A. Saffiotti, “Robot task planning using semantic maps,” *RAS*, 2008.
- [2] C. Huang, O. Mees, A. Zeng, and W. Burgard, “Visual language maps for robot navigation,” in *ICRA*, 2023.
- [3] L. Antanas, P. Moreno, M. Neumann, R. P. de Figueiredo, K. Kersting, J. Santos-Victor, and L. De Raedt, “Semantic and geometric reasoning for robotic grasping: a probabilistic logic approach,” *AR*, 2019.
- [4] W. Shen, G. Yang, A. Yu, J. Wong, L. P. Kaelbling, and P. Isola, “Distilled feature fields enable few-shot language-guided manipulation,” in *CoRL*, 2023.
- [5] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *CVPR*, 2017.
- [6] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *NeurIPS*, 2017.
- [7] Z. Hu, X. Bai, J. Shang, R. Zhang, J. Dong, X. Wang, G. Sun, H. Fu, and C.-L. Tai, “Vmnet: Voxel-mesh network for geodesic-aware 3d semantic segmentation,” in *ICCV*, 2021.
- [8] H.-Y. Meng, L. Gao, Y.-K. Lai, and D. Manocha, “Vv-net: Voxel vae net with group convolutions for point cloud segmentation,” in *ICCV*, 2019.
- [9] R. Hanocka, A. Hertz, N. Fish, R. Giryes, S. Fleishman, and D. Cohen-Or, “Meshcnn: a network with an edge,” *ToG*, 2019.
- [10] S. Zhi, T. Laidlow, S. Leutenegger, and A. J. Davison, “In-place scene labelling and understanding with implicit scene representation,” in *ICCV*, 2021.
- [11] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, “Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving,” *arXiv preprint*, 2023.
- [12] Y. Zhang, Z. Zhu, and D. Du, “Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction,” *arXiv preprint*, 2023.
- [13] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, “Tri-perspective view for vision-based 3d semantic occupancy prediction,” in *CVPR*, 2023.
- [14] 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*, 2021.
- [15] A. Kundu, K. Genova, X. Yin, A. Fathi, C. Pantofaru, L. J. Guibas, A. Tagliasacchi, F. Dellaert, and T. Funkhouser, “Panoptic neural fields: A semantic object-aware neural scene representation,” in *CVPR*, 2022.
- [16] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, “Lerf: Language embedded radiance fields,” *arXiv preprint*, 2023.
- [17] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, “Matterport3d: Learning from rgbd data in indoor environments,” in *3DV*, 2017.
- [18] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, “Occupancy networks: Learning 3d reconstruction in function space,” in *CVPR*, 2019.
- [19] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger, “Convolutional occupancy networks,” in *ECCV*, 2020.
- [20] M. Oechsle, S. Peng, and A. Geiger, “Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction,” in *ICCV*, 2021.
- [21] W. Yang, G. Chen, C. Chen, Z. Chen, and K.-Y. K. Wong, “S³-nerf: Neural reflectance field from shading and shadow under a single viewpoint,” in *NeurIPS*, 2022.
- [22] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, “Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction,” in *NeurIPS*, 2021.
- [23] L. Yaniv, J. Gu, Y. Kasten, and Y. Lipman, “Volume rendering of neural implicit surfaces,” in *NeurIPS*, 2021.
- [24] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger, “Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction,” in *NeurIPS*, 2022.
- [25] D. Azinović, R. Martin-Brualla, D. B. Goldman, M. Nießner, and J. Thies, “Neural rgbd surface reconstruction,” in *CVPR*, 2022.
- [26] J. Wang, T. Bleja, and L. Agapito, “Go-surf: Neural feature grid optimization for fast, high-fidelity rgbd surface reconstruction,” in *3DV*, 2022.
- [27] W. Dong, C. Choy, C. Loop, O. Litany, Y. Zhu, and A. Anandkumar, “Fast monocular scene reconstruction with global-sparse local-dense grids,” in *CVPR*, 2023.
- [28] Y. Xiao, Y. Zhao, Y. Xu, and S. Gao, “Resnerf: Geometry-guided residual neural radiance field for indoor scene novel view synthesis,” *arXiv preprint*, 2022.
- [29] H. Guo, S. Peng, H. Lin, Q. Wang, G. Zhang, H. Bao, and X. Zhou, “Neural 3d scene reconstruction with the manhattan-world assumption,” in *CVPR*, 2022.
- [30] X. Lyu, P. Dai, Z. Li, D. Yan, Y. Lin, Y. Peng, and X. Qi, “Learning a room with the occ-sdf hybrid: Signed distance function mingled with occupancy aids scene representation,” *arXiv preprint*, 2023.
- [31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *ICML*, 2021.
- [32] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, “Scaling up visual and vision-language representation learning with noisy text supervision,” in *ICML*, 2021.
- [33] Q. Liu, Y. Wen, J. Han, C. Xu, H. Xu, and X. Liang, “Open-world semantic segmentation via contrasting and clustering vision-language embedding,” in *ECCV*, 2022.
- [34] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, “Language-driven semantic segmentation,” in *ICLR*, 2022.
- [35] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, “Scaling open-vocabulary image segmentation with image-level labels,” in *ECCV*, 2022.
- [36] J. Xu, J. Hou, Y. Zhang, R. Feng, Y. Wang, Y. Qiao, and W. Xie, “Learning open-vocabulary semantic segmentation models from natural language supervision,” in *CVPR*, 2023.
- [37] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, “Open-vocabulary semantic segmentation with mask-adapted clip,” in *CVPR*, 2023.
- [38] Y.-H. Wu, Y. Liu, X. Zhan, and M.-M. Cheng, “P2t: Pyramid pooling transformer for scene understanding,” *IEEE TPAMI*, 2022.
- [39] W. Zhou, Y. Yue, M. Fang, X. Qian, R. Yang, and L. Yu, “Bcinet: Bilateral cross-modal interaction network for indoor scene understanding in rgbd images,” *Information Fusion*, 2023.
- [40] N. M. M. Shaifullah, C. Paxton, L. Pinto, S. Chintala, and A. Szlam, “Clip-fields: Weakly supervised semantic fields for robotic memory,” in *CoRL workshops*, 2022.
- [41] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser *et al.*, “Openscene: 3d scene understanding with open vocabularies,” in *CVPR*, 2023.
- [42] N. Tsagkas, O. Mac Aodha, and C. X. Lu, “Vi-fields: Towards language-grounded neural implicit spatial representations,” *arXiv preprint*, 2023.
- [43] S. Thrun, “Probabilistic robotics,” *Communications of the ACM*, 2002.
- [44] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma *et al.*, “The replica dataset: A digital replica of indoor spaces,” *arXiv preprint*, 2019.
- [45] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “Scannet: Richly-annotated 3d reconstructions of indoor scenes,” in *CVPR*, 2017.
- [46] D. Rozenberszki, O. Litany, and A. Dai, “Language-grounded indoor 3d semantic segmentation in the wild,” in *ECCV*, 2022.
- [47] Z. Murez, T. van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, and A. Rabinovich, “Atlas: End-to-end 3d scene reconstruction from posed images,” in *ECCV*, 2020.
- [48] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, “Open-vocabulary object detection via vision and language knowledge distillation,” in *ICLR*, 2022.
- [49] J. Lambert, Z. Liu, O. Sener, J. Hays, and V. Koltun, “Mseg: A composite dataset for multi-domain semantic segmentation,” in *CVPR*, 2020.
- [50] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, “Pixel-wise View Selection for Unstructured Multi-View Stereo,” in *ECCV*, 2016.
- [51] J. Wang, P. Wang, X. Long, C. Theobalt, T. Komura, L. Liu, and W. Wang, “Neuris: Neural reconstruction of indoor scenes using normal priors,” in *ECCV*, 2022.