GOV-NeSF: Generalizable Open-Vocabulary Neural Semantic Fields

Yunsong Wang Hanlin Chen Gim Hee Lee
Department of Computer Science, National University of Singapore
{yunsong, hanlin.chen, gimhee.lee}@comp.nus.edu.sg
https://github.com/wangys16/GOV-NeSF

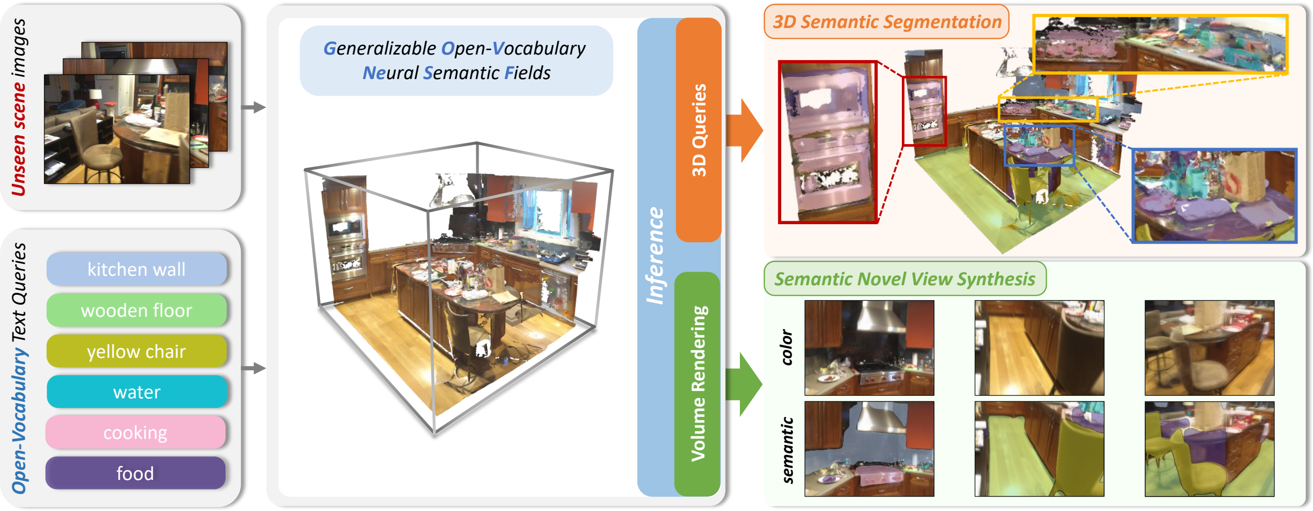

Figure 1. **Overall pipeline of GOV-NeSF.** Given the posed images from any unseen 3D scene, and arbitrary open-vocabulary text queries, our model is capable of both open-vocabulary 3D semantic segmentation and novel view synthesis with 2D semantic segmentation.

Abstract

*Recent advancements in vision-language foundation models have significantly enhanced open-vocabulary 3D scene understanding. However, the generalizability of existing methods is constrained due to their framework designs and their reliance on 3D data. We address this limitation by introducing **G**eneralizable **O**pen-**V**ocabulary **N**eural **S**emantic **F**ields (GOV-NeSF), a novel approach offering a generalizable implicit representation of 3D scenes with open-vocabulary semantics. We aggregate the geometry-aware features using a cost volume, and propose a Multi-view Joint Fusion module to aggregate multi-view features through a cross-view attention mechanism, which effectively predicts view-specific blending weights for both colors and open-vocabulary features. Remarkably, our GOV-NeSF exhibits state-of-the-art performance in both 2D and 3D open-vocabulary semantic segmentation, eliminating the need for ground truth semantic labels or depth priors, and effectively generalize across scenes and datasets without fine-tuning.*

1. Introduction

Semantic segmentation for 2D [13, 25, 34, 45] and 3D [4, 29–31] is a fundamental problem in computer vision with broad applications in fields such as autonomous driving [20, 39], robotic navigation [54], medical imaging analysis [15], *etc*. Given the input 2D images or 3D data such as point cloud, the model is trained to predict dense semantic labels that are assigned to each pixel or point, respectively. Conventional approaches for semantic segmentation are limited by a predefined set of classes that can potentially be assigned to a pixel/point, which hinders the generalizability of the models across datasets and label sets, resulting in dataset-specific models that heavily rely on labels.

Recently, Vision-Language Models (VLMs) [32, 33, 49, 51] propose to learn vision-language correlations from web-scale image-text pairs, showing impressive generalizability in zero-shot vision recognition tasks across different datasets. Attempts like MaskCLIP [56] and PointCLIP [57] have explored transferring knowledge from 2D VLMs to 3D encoders, benefiting from the robustness and generalizability inherent in 2D VLMs. Despite their advantages, these methods typically require pairs of images and point clouds during training, which is largely constrained by the limited availability of 3D data. Additionally, when this knowledge is directly distilled from 2D to 3D contexts, the limited 3D dataset sizes for training may compromise the generalizability of the models, which is the key strength of VLMs.

OpenScene [28] has achieved notable success in zero-shot and open-vocabulary 3D semantic segmentation by averaging multi-view open-vocabulary features and distilling them into a 3D encoder. Despite these achievements, there are several issues with this approach: 1) Averaging multi-view open-vocabulary features can result in sub-optimal performance (*cf.* Table 2); 2) The direct distillation of open-vocabulary features from 2D to 3D can impair the generalizability (*cf.* Table 2); 3) The performance of OpenScene-2D significantly deteriorates during inference without depth maps (*cf.* Table 2, Figure 5). To address these limitations, we leverage Neural Radiance Field (NeRF) to simultaneously encode 3D scene representations and open-vocabulary semantics, where we train the model to learn the blending weights of multi-view open-vocabulary features using the supervision from novel views. While previous works like LERF [16], VL-Fields [37], and Open-NeRF [52] have explored using neural implicit fields to learn open-vocabulary semantics, they still require per-scene optimization and cannot generalize to unseen scenes.

In this paper, we propose **G**eneralizable **O**pen-**V**ocabulary **Ne**ural **S**emantic **F**ields (GOV-NeSF), the overall pipeline of which is shown in Figure 1. GOV-NeSF is trained using only 2D data without the need for point clouds, ground truth semantic labels or depth maps, and can generalize to unseen scenes for open-vocabulary semantic segmentation. Our model is capable of 2D semantic segmentation from novel views and 3D semantic segmentation of the entire 3D scene. Our approach begins with the construction of a cost volume through back-projection of image features, which is then processed by a 3D U-Net [5, 19] to extract geometry-aware features of the 3D scene. Subsequently, during volume rendering, we predict the color and semantic of the sampled 3D points through proposing a Multi-view Joint Fusion Module, which is trained to blend both the color and open-vocabulary values from multi-view projections. Additionally, a Cross-View Attention module is introduced to effectively aggregate multi-view image features before the prediction of blending weights. Extensive experiments validate our state-of-the-art performance on both 2D and 3D open-vocabulary semantic segmentation. Our **contributions** can be summarized as:

1. To the best of our knowledge, we are the first to explore Generalizable Open-Vocabulary Neural Semantic Fields. Its robust design allows for direct inference in unseen scenes and seamless adaptation across datasets.
2. The Multi-view Joint Fusion module, a key innovation of our model, blends colors and open-vocabulary features from multi-view inputs. It employs implicit scene representation to predict geometry-aware blending weights and integrates a cross-view attention module for enhanced multi-view feature aggregation.
3. Extensive experiments demonstrate our state-of-the-art open-vocabulary semantic segmentation results with remarkable generalizability across scenes and datasets.

2. Related Work

Generalizable NeRF. The field of neural implicit representation has seen significant advances [1, 26, 27, 41], yet these methods depend on computationally intensive per-scene optimization. To overcome this limitation, several recent methods have been proposed to place emphasis on the generalization to unseen scenes. Particularly, MVS-NeRF [2], IBRNet [43], Point-NeRF [47], and PixelNeRF [50] are designed to acquire neural radiance fields using images from arbitrary unseen scenes, achieving novel view synthesis without per-scene optimization. Notably, PixelNeRF [50] and IBRNet [43] employ volume rendering techniques through back-projecting features from nearby reference images into 3D space. Instead of inputting only the nearby views of the source view, we feed our model with the coarsely captured images from the entire unseen scene, in order to simultaneously encode the representation and the open-vocabulary semantics of the whole 3D scene.

Neural Semantic Fields. As introduced in [55], Semantic-NeRF marked a significant milestone in neural implicit representation field by incorporating semantics into the NeRF framework. Subsequent researchers have expanded upon this foundational concept. For example, several studies [11, 17, 18] extended Semantic-NeRF by incorporating instance-level modeling, and [38] introduced abstract visual features for post hoc semantic segmentation derivation. [48] presented the concept of an object-compositional neural radiance field. Panoptic NeRF [11] is designed for panoptic radiance fields to address tasks such as label transfer and scene editing, and GNeSF [3] introduces a generalizable neural semantic field using a soft voting mechanism. Recently, people also explored distilling the Vision-Language Models knowledge into neural implicit representation to achieve open-vocabulary neural semantic fields. LERF [16], VL-Fields [37] and Open-NeRF [52] leverage vision-language models to encode the open-vocabulary semantics using neural implicit fields, performing semantic novel view synthesis given arbitrary text query input. Nonetheless, most existing works require per-scene optimization and cannot generalize to unseen scenes. In contrast to the existing methods, we focus on the development of generalizable open-vocabulary neural semantic fields to

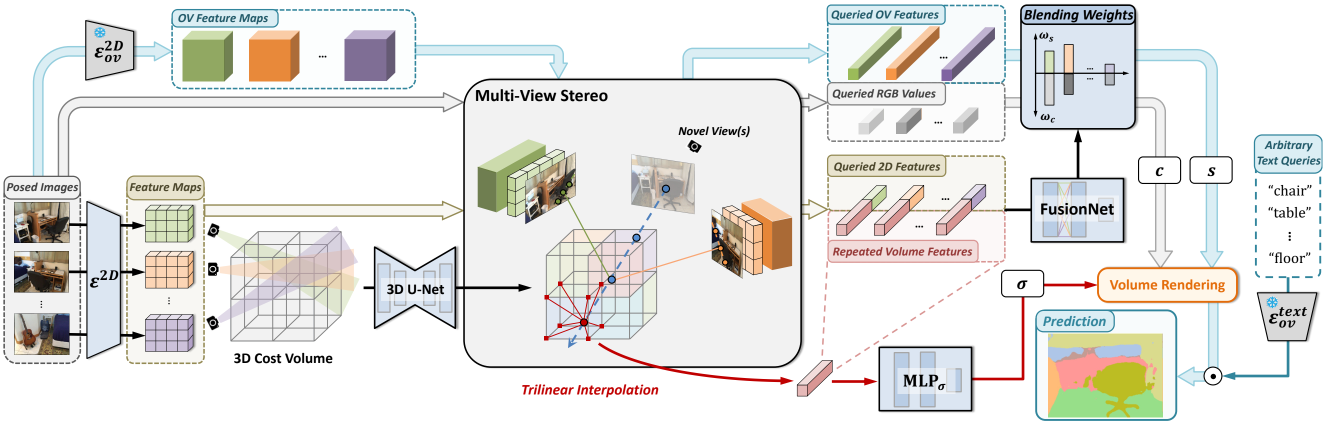

Figure 2. **Structure of GOV-NeSF.** Given a set of posed images of the 3D scene, we first use a shared image encoder to extract the 2D feature maps, and unproject them to build a 3D cost volume. Moreover, we leverage LSeg [21] to predict the per-pixel open-vocabulary features. We then perform Multi-View Stereo to query the 2D and open-vocabulary features for each sampled 3D point along the ray, concatenate the queried 2D features and the volume feature, and feed them into the FusionNet to predict blending weights. The final color and open-vocabulary feature are the weighted sum of multi views using the blending weights.

equip neural semantic fields with the capability to recognize open-world categories across unseen scenes.

Open-Vocabulary 3D Semantic Segmentation. With rapid advancements of Vision-Language Models [8, 12, 21, 22, 32], several works have proposed to distill the open-vocabulary semantic knowledge into 3D models. One line of works [9, 14, 42, 57] distill image-level CLIP features into the 3D semantic segmentation models, which however can suffer from coarse supervision. Recently, OpenScene [28] proposes to leverage LSeg [21], OpenSeg [12] to fuse the dense pixel-wise open-vocabulary features into the point clouds, and achieves remarkable zero-shot and open-vocabulary 3D semantic segmentation results comparing to existing methods. However, OpenScene requires point clouds input during training and inference, and their fusion of point-wise features is naive averaging. Moreover, directly distilling the 2D open-vocabulary features into 3D models can lead to limited generalizability, since the 3D model is typically trained on 3D datasets that are remarkably small comparing to 2D web-scale datasets. In contrast, we propose to train a neural semantic fields, in order to learn the multi-view features blending instead of direct distillation, thus further unleashing the generalizability of 2D VLMs.

3. Methodology

3.1. Overview

The overall framework of our proposed GOV-NeSF is shown in Figure 2. Given posed images of a 3D scene, we first use an off-the-shelf 2D Open-Vocabulary semantic segmentation model LSeg [21] to extract the per-pixel open-vocabulary feature maps. We also train an image encoder to embed image features, which is followed by back-projection to build a 3D cost volume. Subsequently, during volume rendering, we leverage Multi-View Stereo (MVS) to query features for each sampled 3D point, and propose a FusionNet to blend the multi-view colors and open-vocabulary features with cross-view attention.

3.2. Feature Extraction

Feature Fusion. Formally, given a set of posed images $\{\boldsymbol{I}^n\}_{n=1}^N$, $\boldsymbol{I} \in \mathbb{R}^{3\times h\times w}$, we train an image encoder $\boldsymbol{\epsilon}^{2D}$ to extract the 2D feature maps $\{\boldsymbol{F}^n\}_{n=1}^N$, $\boldsymbol{F}^n \in \mathbb{R}^{c\times\bar{h}\times\bar{w}}$, and build a 3D Cost Volume through unprojection. Specifically, for a cost volume $\boldsymbol{V}_c \in \mathbb{R}^{C\times H\times W\times D}$, each voxel feature is accumulated as:

$$\boldsymbol{v}[:, i, j, k] = [\mathcal{A}_{i,j,k}, \mathcal{V}_{i,j,k}], \tag{1}$$

where [ ] denotes concatenation, $\mathcal{A}$, $\mathcal{V}$ represent average and variance of the features that are unprojected to the voxel. We then use a 3D U-Net [5, 19] to extract the geometry features of the 3D scene and derive the aggregated Volume Features $\boldsymbol{V}_a \in \mathbb{R}^{C'\times H\times W\times D}$.

Open-Vocabulary Feature Extraction. To leverage 2D vision-language models to provide dense open-vocabulary features, we use a pre-trained LSeg [21] model to predict Open-Vocabulary (OV) feature maps $\{\boldsymbol{F}^n_{ov}\}_{n=1}^N$, where $\boldsymbol{F}^n_{ov} \in \mathbb{R}^{d\times h\times w}$ share the same image height h and width w as $\boldsymbol{I}^n$, with per-pixel d-dimensional OV feature.

3.3. Multi-view Joint Fusion

Multi-View Stereo. As shown in Figure 3, given the input posed images $\{\boldsymbol{I}^n\}_{n=1}^N$ as reference images, and their 2D feature maps $\{\boldsymbol{F}^n\}_{n=1}^N$ and OV feature maps $\{\boldsymbol{F}^n_{ov}\}_{n=1}^N$,

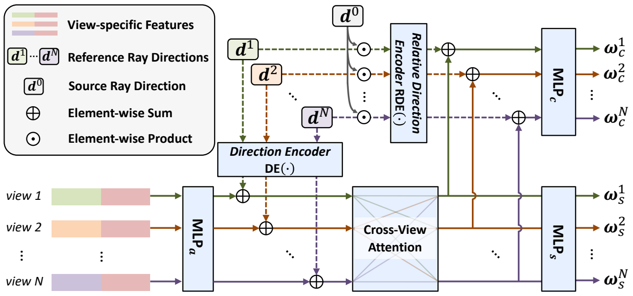

Figure 3. **FusionNet Structure.** We aggregate multi-view features through Cross-View Attention module, and predict view-specific blending weights. Refer to the text for more details.

we perform volume rendering to render images from novel views. Given a ray $\{\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d}, t \geq 0\}$ from the source view, we project a 3D point $\boldsymbol{p}$ sampled on the ray onto the reference views and get its normalized 2D coordinates $\{\pi_n(\boldsymbol{p})\}_{n=1}^{N}$. We then use bilinear interpolation to get the queried RGB values $\{\boldsymbol{i}^n = \boldsymbol{I}^n(\pi_n(\boldsymbol{p}))\}_{n=1}^{N}$, 2D features $\{\boldsymbol{f}^n = \boldsymbol{F}^n(\pi_n(\boldsymbol{p}))\}_{n=1}^{N}$, and OV feature $\{\boldsymbol{f}_{ov}^n = \boldsymbol{F}_{ov}^n(\pi_n(\boldsymbol{p}))\}_{n=1}^{N}$. Furthermore, we query the volume feature at $\boldsymbol{p}$ using trilinear interpolation to get $\boldsymbol{v} = \boldsymbol{V}_a(:, \boldsymbol{p})$.

Joint View Blending. Since we are targeting generalizable neural implicit fields for room-scale scene representation, it is difficult to directly regress colors and OV features in the 3D space (*cf*. Table 1, Table 2, Table 3). We thus propose to jointly blend the colors and the OV features from reference views instead of direct regression. Specifically, we construct the view-specific features for n-th view as $\boldsymbol{x}^n = [\boldsymbol{f}^n, \boldsymbol{v}]$, and feed $\{\boldsymbol{x}^n\}_{n=1}^{N}$ into our proposed FusionNet to predict view-specific weights:

$$\{\boldsymbol{w}_c^n\}_{n=1}^{N}, \{\boldsymbol{w}_s^n\}_{n=1}^{N} = \text{FusionNet}(\{\boldsymbol{x}^n\}_{n=1}^{N}), \quad (2)$$

where $\{\boldsymbol{w}_c^n\}_{n=1}^{N}$, $\{\boldsymbol{w}_s^n\}_{n=1}^{N}$ are the blending weights for colors and OV features, respectively. We then normalize them using Softmax to get $\{\bar{\boldsymbol{w}}_c^n\}_{n=1}^{N}$ and $\{\bar{\boldsymbol{w}}_s^n\}_{n=1}^{N}$, and compute the color $\boldsymbol{c}$ and OV feature $\boldsymbol{s}$ of 3D point $\boldsymbol{p}$ as:

$$\boldsymbol{c} = \sum_{n=1}^{N} \bar{\boldsymbol{w}}_c^n \cdot \boldsymbol{i}^n, \quad \boldsymbol{s} = \sum_{n=1}^{N} \bar{\boldsymbol{w}}_s^n \cdot \boldsymbol{f}_{ov}^n. \quad (3)$$

Additionally, we predict the density of $\boldsymbol{p}$ from the volume feature: $\sigma = \text{MLP}_\sigma(\boldsymbol{v})$.

FusionNet. As shown in Figure 3, given the cross-view features $\{\boldsymbol{x}^n\}_{n=1}^{N}$, we first use a shared tiny MLP to aggregate features for each view: $\{\tilde{\boldsymbol{x}}^n\}_{n=1}^{N} = \text{MLP}_a(\{\boldsymbol{x}^n\}_{n=1}^{N})$. Subsequently, we propose a Cross-View Attention (CVA) module based on self-attention [40] to exchange information across views before blending weights prediction:

$$\{\tilde{\boldsymbol{x}}_{att}^n\}_{n=1}^{N} = \text{CVA}\left(\{\tilde{\boldsymbol{x}}^n + \text{DE}\left(\boldsymbol{d}^n\right)\}_{n=1}^{N}\right), \quad (4)$$

where CVA(·) is a self-attention module based on Transformers [10, 40], $\boldsymbol{d}^n$ is the ray direction from n-th camera to $\boldsymbol{p}$, and DE(·) is the ray direction encoder.

To equip the color prediction with source-view-dependency, we compute the color blending weights using:

$$\{\boldsymbol{w}_c^n\}_{n=1}^{N} = \text{MLP}_c\left(\{\tilde{\boldsymbol{x}}_{att}^n + \text{RDE}\left(\boldsymbol{d}^0 \cdot \boldsymbol{d}^n\right)\}_{n=1}^{N}\right), \quad (5)$$

where $\boldsymbol{d}^0$ is the ray direction from source camera to $\boldsymbol{p}$, and RDE(·) is relative ray direction encoder. Furthermore, the prediction of OV feature blending weights should be source-view-agnostic:

$$\{\boldsymbol{w}_s^n\}_{n=1}^{N} = \text{MLP}_s\left(\{\tilde{\boldsymbol{x}}_{att}^n\}_{n=1}^{N}\right). \quad (6)$$

Joint Volume Rendering. Similarly to volume rendering [26], we accumulate both the colors and OV features along each ray using the shared density field:

$$\hat{\boldsymbol{C}}(\boldsymbol{r}) = \sum_{i=1}^{N_p} T_i \left(1 - \exp(-\sigma_i \delta_i)\right) \boldsymbol{c}_i, \quad (7a)$$

$$\hat{\boldsymbol{S}}(\boldsymbol{r}) = \sum_{i=1}^{N_p} T_i \left(1 - \exp(-\sigma_i \delta_i)\right) \boldsymbol{s}_i, \quad (7b)$$

$$T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right), \quad (7c)$$

where δ_j is the distance between sampled points along the ray. $\{\boldsymbol{c}_i\}_{i=1}^{N_p}$, $\{\boldsymbol{s}_i\}_{i=1}^{N_p}$, $\{\sigma_i\}_{i=1}^{N_p}$ are the colors, OV features, and densities of the points along the ray.

3.4. Training Objective

Our loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{color} + \alpha \mathcal{L}_{ov}, \quad (8)$$

where α is the weight balancing two loss terms. The color loss $\mathcal{L}_{color}$ and open-vocabulary feature loss $\mathcal{L}_{ov}$ are respectively defined as:

$$\mathcal{L}_{color} = \frac{1}{|\mathcal{R}|} \sum_{\boldsymbol{r} \in \mathcal{R}} \left\| \hat{\boldsymbol{C}}(\boldsymbol{r}) - \boldsymbol{C}(\boldsymbol{r}) \right\|_2^2, \quad (9a)$$

$$\mathcal{L}_{ov} = \frac{1}{|\mathcal{R}|} \sum_{\boldsymbol{r} \in \mathcal{R}} \left(1 - \cos\left(\hat{\boldsymbol{S}}(\boldsymbol{r}), \boldsymbol{S}(\boldsymbol{r})\right)\right), \quad (9b)$$

where $\boldsymbol{C}(\boldsymbol{r})$ and $\boldsymbol{S}(\boldsymbol{r})$ are the ground truth color and OV feature for ray $\boldsymbol{r}$, and cos(·) is cosine similarity. Note that since the OV feature map prediction from LSeg is not view-consistent, we detach the gradient from $\mathcal{L}_{ov}$ to $\boldsymbol{\sigma}$ to enhance the quality of the learned density fields.

3.5. Inference

Given input posed images of a 3D scene that is unseen during training, we first encode them into our framework to build the scene representation with OV semantics. We then input an arbitrary set of texts $\{\boldsymbol{t}^k\}_{k=1}^K$ and encode the text features using the frozen text encoder $\boldsymbol{\epsilon}_{ov}^{text}$: $\{\boldsymbol{\epsilon}_{ov}^{text}(\boldsymbol{t}^k)\}_{k=1}^K$, which are utilized for both 2D and 3D Open-Vocabulary Semantic Segmentation.

2D semantic segmentation. We render the OV feature maps from arbitrary novel views: $\{\hat{\boldsymbol{F}}_{ov}^m\}_{m=1}^M$ and compute its pixel-wise cosine similarity with $\{\boldsymbol{f}_t(\boldsymbol{t}^k)\}_{k=1}^K$. The per-pixel segmentation result is then given by the argmax within the similarities with all the query text features.

3D semantic segmentation. Our model is also capable of segmenting any given point clouds corresponding to the input images. We simply query the source-view-agnostic OV features of the given point coordinates and perform per-point segmentation similarly as 2D semantic segmentation. Note that comparing to OpenScene-2D [28] which uses naive averaging between multi-view OV feature maps, our model leverages neural implicit representation to learn the blending weights based on the supervision from novel views OV feature maps.

Remarks. Although our method does not require the ground truth depth maps during training, we can also leverage the depth maps during inference similarly as OpenScene-2D [28] through a Depth Guided Masking (DGM) method. We compare the distances from 3D points to multi views, and mask out the blending weights of the views where the distances differ over 25% from ground truth depths. In the subsequent experiments, "Ours" does not include the DGM module unless otherwise specified.

4. Experiments

We conduct extensive experiments on 2D and 3D generalizable open-vocabulary semantic segmentation tasks. We also compare our novel view synthesis performance with existing generalizable NeRF methods to quantify our scene representation quality. Furthermore, we perform detailed qualitative comparisions on both 2D and 3D open-vocabulary semantic segmentation, and conduct extensive ablation studies on our proposed components.

4.1. Experiment Setup

Implementation Details. For 2D feature extraction, we leverage ResUNet-34 [7] to encode 2D feature maps, and use the ViT-L/16-based LSeg [21] model to extract OV feature maps. For 3D feature aggregation, we leverage a 3D ResUNet [19] containing 3 levels. For volume rendering, we sample 64 points along each ray and do not perform hierarchical sampling to reduce GPU consumption.

Datasets. We mainly conduct experiments on two datasets: real-world dataset ScanNet [6], and synthetic dataset Replica [35]. ScanNet is a large RGB-D dataset containing 2.5M views in 1,513 real-world 3D indoor scenes with the corresponding camera poses and semantic labels. We train our models on their provided training set split and evaluate on the validation set. Replica provides 18 high-quality synthetic 3D indoor scenes with mesh and ground truth 2D semantic labels. We follow Semantic-NeRF [55] to generate sets of posed images in 8 scenes. Since Replica [35] does not provide the ground truth 3D semantic labels, we leverage the TSDF Fusion in [36] to generate the mesh with semantic labels. To evaluate the generalizability of our proposed model, we train our models on ScanNet train set and evaluate them on *both* ScanNet val set and Replica. For each scene, we extract 100 images with corresponding camera poses. During training, we input 30 posed images as reference images and render 3 images from novel views for each iteration. During testing, we input 95 images and render 5 images from novel views for evaluating 2D semantic segmentation, and input all 100 images for evaluating 3D semantic segmentation.

Metrics. To evaluate the semantic segmentation quality, we compute mean Intersection-over-Union (mIoU), total Accuracy (oAcc), and average Accuracy (mAcc) on both 2D and 3D semantic segmentation. Similar to previous works [26, 43], we report PSNR, SSIM [44] and LPIPS$_{vgg}$ [53] to evaluate the quality of our novel view synthesis.

Baselines. We build the 2D and 3D baselines since there is no prior work on generalizable open-vocabulary neural semantic fields:

- **2D Semantic Segmentation.** We modify the model of S-Ray [23] to S-Ray-OV, which predicts open-vocabulary feature maps instead of one-hot semantic logits to reproduce their open-vocabulary results. We also compare with the Distill Baseline, which shares the basic framework design as ours without the Multi-view Joint Fusion module, *i.e.* directly regressing open-vocabulary features using the multi-view image features and volume features.
- **3D Semantic Segmentation.** We mainly compare with the Distill Baseline, and LSeg [21]-based OpenScene-2D and OpenScene-3D [28]. For a comprehensive analysis, we compare with OpenScene-2D in both *w/* and *w/o* Depth scenarios. For evaluating OpenScene-3D on Replica, we use their provided model pre-trained on ScanNet with fused LSeg features. We first input their preprocessed point clouds of Repilca scenes to predict the dense OV features, and transfer them to the meshes that we extracted using TSDF Fusion.

Furthermore, we still compare with NeRF-Det [46], NeuRay [24] and S-Ray [23] in terms of novel view synthesis to evaluate the quality of our 3D scene representations despite novel view synthesis is not the focus of GOV-NeSF.

	Method	ScanNet [6]			Replica [35]		
		mIoU	oAcc	mAcc	mIoU	oAcc	mAcc
Fully-Supervised	MVSNeRF [2]+Semantic Head†	39.8	60.0	46.0	23.4	54.3	33.7
	NeuRay [24]+Semantic Head†	51.0	77.6	57.1	35.9	69.4	44.0
	S-Ray [23]†	57.2	78.2	62.6	41.6	70.5	47.2
Open-Vocabulary	S-Ray [23]-OV	33.9	50.6	45.7	9.7	26.7	18.2
	Distill Baseline	46.4	69.0	56.5	5.0	28.1	11.5
	Ours	**52.2**	**73.8**	**62.2**	**44.3**	**76.2**	**57.6**
	LSeg$_{rd}$ [21]‡	48.4	69.4	57.6	23.2	53.7	31.3
	LSeg$_{gt}$ [21]‡	55.9	77.3	65.4	52.0	79.6	64.9

Table 1. **Generalizable NeRF-based 2D Semantic Segmentation Results.** We report semantic segmentation results from novel views in novel scenes. † are results reported in [23]. ‡ LSeg$_{rd}$, LSeg$_{gt}$ are the results of directly applying LSeg [21] on the rendered images by NeuRay [24] or ground truth images at novel views, respectively.

Model Input	Method	ScanNet [6]			Replica [35]		
		3D mIoU	3D oAcc	3D mAcc	3D mIoU	3D oAcc	3D mAcc
3D	OpenScene-3D [28]†	51.6	72.8	62.5	3.3	27.7	6.5
2D *w/o Depth*	OpenScene-2D [28] *w/o* Depth†	42.1	64.4	56.8	31.2	64.4	48.7
	Distill Baseline	43.4	64.0	56.2	17.7	42.7	31.0
	Ours	**45.7**	**68.3**	**60.3**	**32.8**	**66.3**	**49.4**
2D *w/ Depth*	OpenScene-2D [28]†	52.2	73.4	63.3	35.8	69.4	52.2
	Ours *w/* DGM	**53.5**	**74.6**	**64.4**	**37.7**	**71.5**	**53.3**

Table 2. **Open-Vocabulary 3D Semantic Segmentation Results.** † are our reproduced results.

Method	ScanNet [6]			Replica [35]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRF-Det [46]†	18.8	0.743	0.505	13.5	0.664	0.618
NeuRay [24]†	**22.8**	0.786	**0.142**	18.1	0.606	**0.195**
S-Ray [23]†	22.0	0.769	0.154	18.0	0.604	0.203
Ours	21.5	**0.801**	0.456	**21.5**	**0.811**	0.419

Table 3. **Novel View Synthesis Results.** † are our reproduced results using our extracted datasets.

4.2. Main Results

2D Semantic Segmentation. In Table 1, we show semantic segmentation results of various Generalizable NeRF-based methods. Our Distill Baseline surpasses S-Ray [23]-OV and performs comparably to LSeg$_{rd}$ [21] on ScanNet. However, when applied to the Replica dataset, the Distill Baseline shows significantly reduced effectiveness, highlighting the undermined generalizability in direct distillation. In contrast, our method exhibits a substantial improvement on ScanNet, outperforming S-Ray-OV by a margin of $+18.3$ in mIoU and the Distill Baseline by $+5.8$ mIoU. More notably, on the Replica dataset, our approach exceeds both the S-Ray-OV and Distill Baseline by over $+34.6$ mIoU and $+39.4$ in mAcc. These significant improvements can be attributed to our Joint Blending module, which blends multi-view OV features instead of direct regression. Furthermore, our approach outperforms LSeg$_{rd}$ [21] by $+3.8$ mIoU in ScanNet and $+21.1$ mIoU in Replica. Note that our given images are relatively sparse, which results in challenges in rendering high-quality images from novel views, yet our performance remains comparble to LSeg$_{gt}$ when transferred to Replica without fine-tuning.

3D Semantic Segmentation. Table 2 shows the comparisons with different open-vocabulary 3D semantic segmentation methods. OpenScene-3D performs comparably to OpenScene-2D and our method in ScanNet, yet greatly degenerates when transferred to Replica, necessitating the usage of 2D VLMs during inference. Under the *w/o Depth* setting, we reproduce OpenScene-2D [28] results without

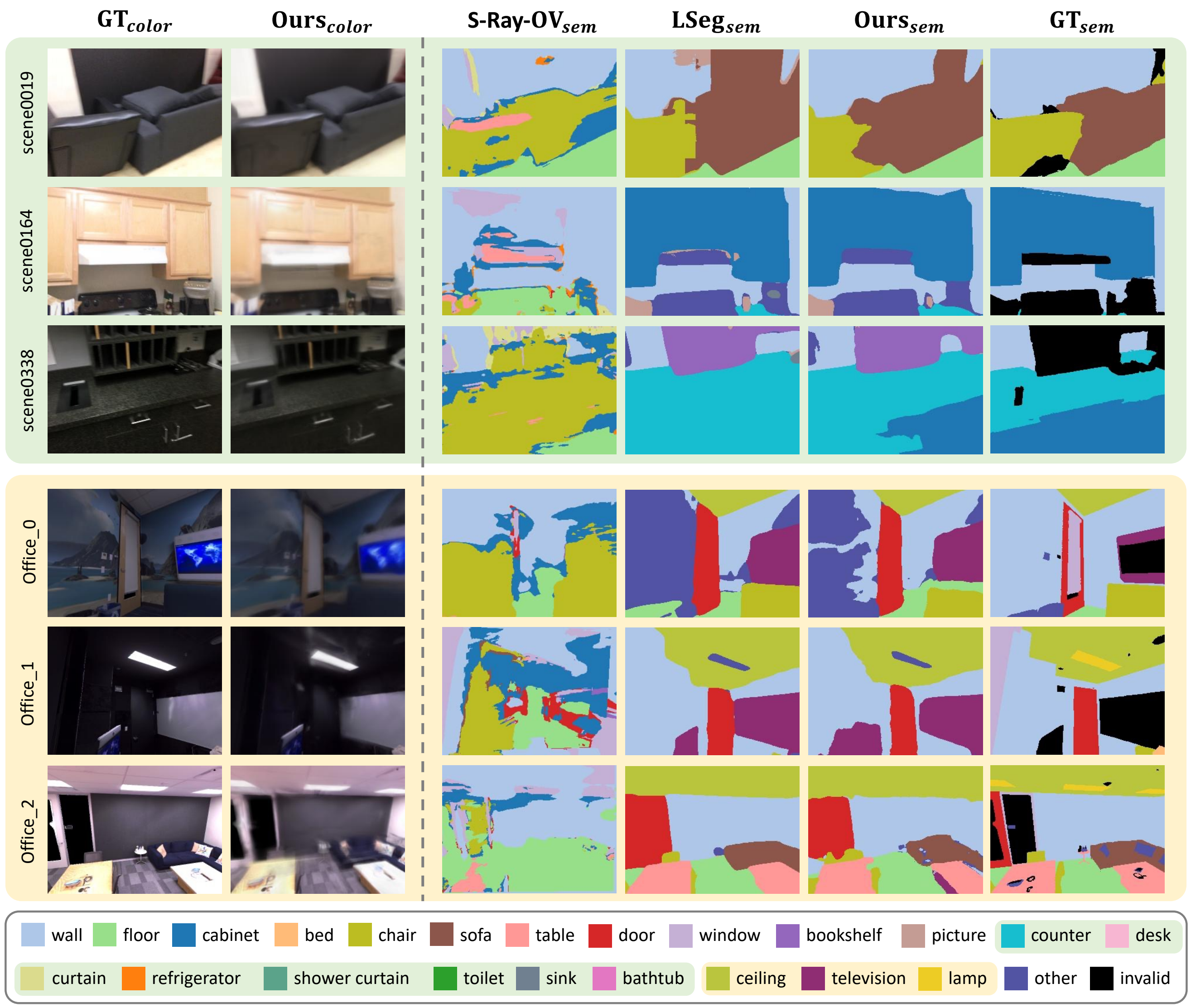

Figure 4. **Visualization of 2D results.** We show the GT color images, our rendered color images, S-Ray [23]-OV rendered semantics, our rendered semantics, LSeg$_{gt}$ [21] predictions, and GT semantics on novel views from unseen scenes in ScanNet [6] and Replica [35].

depth-based masking. Our Distill Baseline performs comparably as the OpenScene-2D *w/o Depth* on ScanNet, and our approach surpasses both baselines by at least $+2.3$ mIoU on ScanNet and $+1.6$ mIoU on Replica. Notably, our approach surpasses OpenScene-3D and Distill Baseline by $+29.5$ mIoU and $+15.1$ mIoU on Replica, respectively, thus further emphasizing the robustness of our model. Additionally, our method can also leverage depth maps during inference when available to further improve OpenScene-2D performance by $+1.3 \sim +1.9$ mIoU.

Novel View Synthesis. We also evaluate the novel view synthesis quality in Table 3, where our approach achieves superior performance than NeRF-Det [46] and comparable results as NeuRay [24] and S-Ray [23]. Note that both NeuRay and S-Ray leverage the ground truth depth maps during training, while we only learn the density field based on RGB images. Moreover, we consistently maintains the quality of scene representation when transferred to Replica.

4.3. Qualitative Results

2D results. We visualize the 2D results in Figure 4, where color images and open-vocabulary semantics are rendered from novel views in unseen scenes. Our color image renderings are relatively blurry compared to the rendering in S-Ray [23] since we are targeting room-scale representation without depth priors. We demonstrate significant improvements over S-Ray [23]-OV since they fail to effectively regress the OV features through direct distillation.

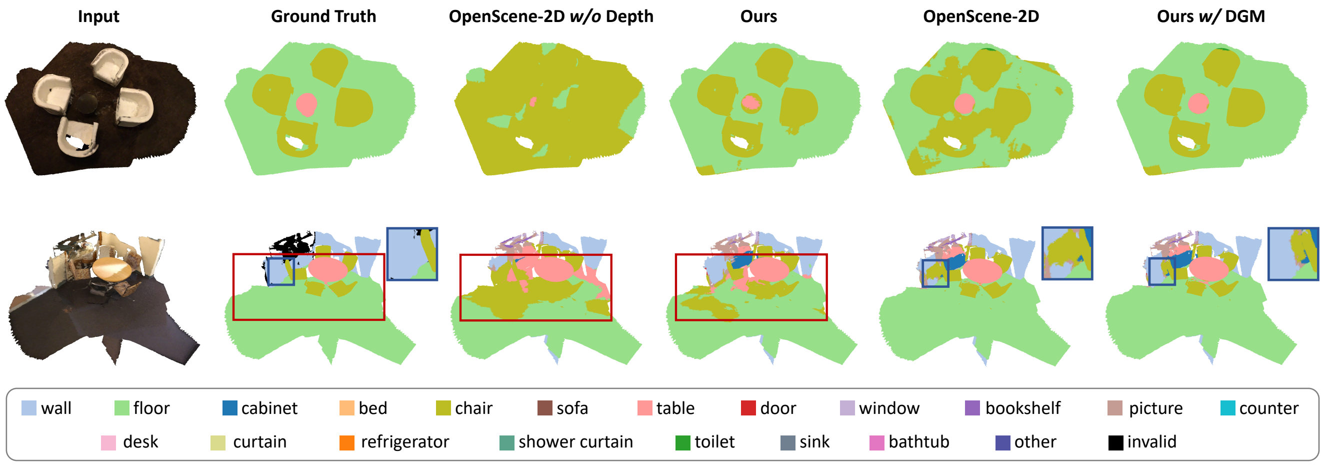

Figure 5. **Visualization of 3D results.** We compare with OpenScene-2D [28] in terms of 3D semantic segmentation on ScanNet.

MJF	VFA	CVA	DGM	2D *Seg.*			3D *Seg.*		
				mIoU	oAcc	mAcc	mIoU	oAcc	mAcc
	✓			46.4	69.0	56.5	39.4	62.2	54.1
✓				49.5	71.8	60.4	44.2	67.1	57.5
✓	✓			50.7	72.1	60.9	44.6	67.4	57.8
✓	✓	✓		**52.2**	**73.8**	**62.2**	45.7	68.3	59.1
✓	✓	✓	✓	46.8	69.8	58.3	**53.5**	**74.6**	**64.2**

Table 4. **Ablation study** on ScanNet, evaluating contributions of our proposed components.

Our results are also comparable to the LSeg$_{gt}$ [21], and in some cases (row 1,3,4) we can surpass their results through the aggregation of multi-view OV features.

3D results. We visualize the 3D results in Figure 5. Our approach demonstrates significant improvements over OpenScene-2D [28] when not given ground truth depth maps, and further refines multi-view OV features fusion when given ground truth depth maps. Our MJF module automatically learns to reason about the occlusion without depth supervision, and facilitates more efficient inference compared to the volume rendering-based depth estimation.

4.4. Ablation Study

Table 4 shows the ablation study we conducted on ScanNet. The first row with the same design as Distill Baseline, greatly suffers from domain gaps and significantly degenerates when transferred to Replica (*cf.* Table 1). This demonstrates the necessity of our MJF module to learn the blending weights for OV features. Furthermore, the integration of Volume Feature Aggregation and Cross-View Attention module improves both 2D and 3D semantic segmentation performance. This improvement underscores the benefits of aggregating scene-level geometry features and cross-view image features prior to blending weights prediction. Moreover, the Depth-Guided Masking module can lead to significant improvements on 3D semantic segmentation by explicitly masking out the occluded projections. However, it also causes a drop in 2D semantic segmentation performance since it can result in empty holes in the rendered images. Consequently, DGM module is only used in 3D semantic segmentation when depth masks are available.

5. Conclusion

In this paper, we introduce GOV-NeSF, a pioneering framework for generalizable open-vocabulary neural semantic fields. Leveraging the neural implicit representation, GOV-NeSF is designed to learn the joint blending weights for both colors and open-vocabulary features queried from multi-view images, eliminating the need for 3D data, depth priors, or explicit ground truth semantic labels. Our framework design enables GOV-NeSF to excel in open-vocabulary semantic segmentation across both 2D semantic NVS and 3D, and set state-of-the-art in benchmarks.

6. Acknowledgement

This work is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 2

[2] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 2, 6

[3] Hanlin Chen, Chen Li, Mengqi Guo, Zhiwen Yan, and Gim Hee Lee. Gnesf: Generalizable neural semantic fields. *arXiv preprint arXiv:2310.15712*, 2023. 2

[4] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 1

[5] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 2, 3

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 5, 6, 7

[7] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020. 5

[8] Jian Ding, Nan Xue, Guisong Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11573–11582, 2021. 3

[9] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7010–7019, 2023. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. *arXiv preprint arXiv:2010.11929*, 2010. 4

[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 *International Conference on 3D Vision (3DV)*, 2022. 2

[12] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[14] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22157–22167, 2023. 3

[15] Feng Jiang, Aleksei Grigorev, Seungmin Rho, Zhihong Tian, YunSheng Fu, Worku Jifara, Khan Adil, and Shaohui Liu. Medical image semantic segmentation based on deep learning. *Neural Computing and Applications*, 29:1257–1265, 2018. 1

[16] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[17] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *Advances in Neural Information Processing Systems*, 2022. 2

[18] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12871–12881, 2022. 2

[19] Kisuk Lee, Jonathan Zung, Peter Li, Viren Jain, and H Sebastian Seung. Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120*, 2017. 2, 3, 5

[20] Baojun Li, Shun Liu, Weichao Xu, and Wei Qiu. Real-time object detection and semantic segmentation for autonomous driving. In *MIPPR 2017: Automatic Target Recognition and Navigation*, pages 167–174. SPIE, 2018. 1

[21] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 3, 5, 6, 7, 8

[22] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 3

[23] Fangfu Liu, Chubin Zhang, Yu Zheng, and Yueqi Duan. Semantic ray: Learning a generalizable semantic field with cross-reprojection attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17386–17396, 2023. 5, 6, 7

[24] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7824–7833, 2022. 5, 6, 7

[25] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the european conference on computer vision (ECCV)*, pages 552–568, 2018. 1

[26] Ben Mildenhall, Pratul Srinivasan, Matthew Tancik, Jonathan Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 2, 4, 5

[27] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2

[28] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–824, 2023. 2, 3, 5, 6, 8

[29] Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Real-time progressive 3d semantic segmentation for indoor scenes. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1089–1098. IEEE, 2019. 1

[30] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[31] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 1

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[35] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 5, 6, 7

[36] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. *CVPR*, 2021. 5

[37] Nikolaos Tsagkas, Oisin Mac Aodha, and Chris Xiaoxuan Lu. Vl-fields: Towards language-grounded neural implicit spatial representations. *arXiv preprint arXiv:2305.12427*, 2023. 2

[38] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3D distillation of self-supervised 2D image representations. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2022. 2

[39] Yu-Ho Tseng and Shau-Shiun Jan. Combination of computer vision detection and segmentation for autonomous driving. In *2018 IEEE/ION Position, Location and Navigation Symposium (PLANS)*, pages 1047–1052. IEEE, 2018. 1

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[41] 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. 2

[42] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 3

[43] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 2, 5

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[45] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 1

[46] Chenfeng Xu, Bichen Wu, Ji Hou, Sam Tsai, Ruilong Li, Jialiang Wang, Wei Zhan, Zijian He, Peter Vajda, Kurt Keutzer, et al. Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23320–23330, 2023. 5, 6, 7

[47] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. *arXiv preprint arXiv:2201.08845*, 2022. 2

[48] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui.

Learning object-compositional neural radiance field for editable scene rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13779–13788, 2021. 2

[49] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 1

[50] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 2

[51] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1

[52] Hao Zhang, Fang Li, and Narendra Ahuja. Open-nerf: Towards open vocabulary nerf decomposition. *arXiv preprint arXiv:2310.16383*, 2023. 2

[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[54] Yuxiao Zhang, Haiqiang Chen, Yiran He, Mao Ye, Xi Cai, and Dan Zhang. Road segmentation for all-day outdoor robot navigation. *Neurocomputing*, 314:316–325, 2018. 1

[55] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 2, 5

[56] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 1

[57] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. Pointclip v2: Adapting clip for powerful 3d open-world learning. *arXiv preprint arXiv:2211.11682*, 2022. 1, 3