Open-Vocabulary 3D Semantic Segmentation with Text-to-Image Diffusion Models

Xiaoyu Zhu[1†], Hao Zhou[2], Pengfei Xing[2], Long Zhao[2], Hao Xu[3], Junwei Liang[4], Alexander Hauptmann[1], Ting Liu[2], and Andrew Gallagher[2]

[1] Carnegie Mellon University
[2] Google DeepMind
[3] Google
[4] HKUST

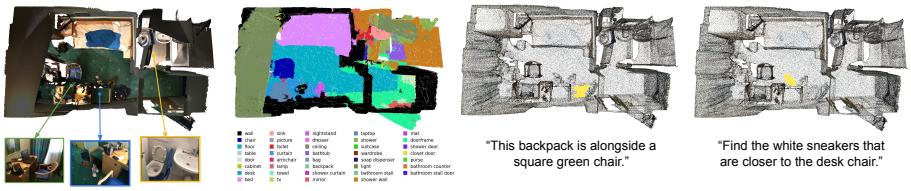

Fig. 1: Illustration of open-vocabulary 3D semantic scene understanding. We propose *Diff2Scene*, a 3D model that performs open-vocabulary semantic segmentation and visual grounding tasks given novel text prompts, without relying on any annotated 3D data. By leveraging discriminative-based and generative-based 2D foundation models, *Diff2Scene* can handle a wide variety of novel text queries for both common and rare classes, like "desk" and "soap dispenser". It can also handle compositional queries, such as "find the white sneakers that are closer to the desk chair."

Abstract. In this paper, we investigate the use of diffusion models which are pre-trained on large-scale image-caption pairs for open-vocabulary 3D semantic understanding. We propose a novel method, namely Diff2Scene, which leverages frozen representations from text-image generative models, along with salient-aware and geometric-aware masks, for open-vocabulary 3D semantic segmentation and visual grounding tasks. Diff2Scene gets rid of any labeled 3D data and effectively identifies objects, appearances, materials, locations and their compositions in 3D scenes. We show that it outperforms competitive baselines and achieves significant improvements over state-of-the-art methods. In particular, Diff2Scene improves the state-of-the-art method on ScanNet200 by 12%.

Keywords: 3D Semantic Understanding · Open-Vocabulary Perception · Diffusion Model

†This work was partially done while the author was a student researcher at Google.

1 Introduction

3D semantic scene understanding, with the task of assigning semantics to every 3D point, plays a fundamental role in many computer vision applications, such as robotics [88], autonomous driving [36], human-computer interaction [22], and augmented reality [27]. Traditional studies in this field usually target solving this problem in a closed-set fashion [16,73], resulting in models that can only be used to make predictions within the predefined label space.

Recent progress in computer vision have witnessed the emerging interests in solving semantic understanding tasks in open-vocabulary settings [35,62,67,78, 94]. In contrast to closed-set setting, models targeting open-vocabulary tasks must make predictions for any semantics described in text, including object category and fine-grained attributes (*e.g.*, shape, color, material, property) as well as their complicated compositions. However, this is a challenging task due to the wide diversity and complexity of possible queries. Motivated by the advance of aligning text and image embeddings with large-scale foundation models [2, 39,48,65], existing methods mitigate this challenge by lifting the image features from foundation models such as CLIP [65] or their descendants [25, 47] to 3D. These lifted feature representations for 3D points can then be used to query with open-vocabulary descriptions, achieving semantic understanding in 3D. Despite these achievements, contrastively trained CLIP-based models exhibit limitations in handling fine-grained classes [66] and novel compositional text queries [58], restricting their performance in open-vocabulary 3D semantic understanding.

The recently developed text-to-image diffusion models have shown outstanding abilities for image generation even with challenging text prompts [55,68,72], such as combinational descriptions with multiple attributes (*e.g., A bucket bag made of blue suede with intricate golden paisley patterns.*) The internal visual representation of these models, entangled with text embedding through cross-attention, have proven correlate well with semantic concepts described by language [43, 60, 61, 89]. On the other hand, CLIP-based foundation models have been shown to struggle with compositionality [58]. Moreover, compared with the CLIP model which is optimized for global representation, diffusion models have proven to be superior at local representation [80], which is a key for dense prediction tasks. Specifically, ODISE [87] applied the internal representations of Stable Diffusion [68] to open-vocabulary 2D semantic understanding tasks and achieved promising results.

One of the key challenges in 3D perception is the severe scarcity of point clouds and their dense labels. Several existing methods have been proposed to solve the lack of data issue in a zero-shot fashion by leveraging the CLIP model pre-trained on large-scale text-image data [37,62,79]. The prior art [62] extracts dense CLIP features from 2D images and distill the knowledge of their lifted 3D counterpart into a 3D mask predictor. However, CLIP features, as discussed above, struggle to handle fine-grained classes [66] and show worse localization capability compared with diffusion features. We leverage diffusion model as feature backbone along with a mask-based segmentation head (*e.g.*, Mask2Fromer [10]) for its intrinsic nature that decouples mask and its semantic representations.

This is intuitively suitable for leveraging semantically-rich embeddings from 2D foundation models and further learning geometrically-accurate masks from the 3D branch. However, performing multi-modal distillation with mask-based segmentation head is a non-trivial task. The frozen features extracted from the decoder of the U-Net in the diffusion model are trained with generative objectives, and cannot be directly used for the perception task. Therefore, directly distilling knowledge from these features as normally done in prior art [54,56,62] is infeasible. Another intuitive way is to leverage a supervised 3D mask proposal network and pool the feature representations from 2D CLIP features for each mask [79]. However, the training of 3D mask proposal network requires labeled 3D data, which may not be feasible in practice.

To mitigate these issues, we propose a novel mask distillation method tailored to distill knowledge from the Mask2Former style 2D branch [10, 87] to the 3D branch, which is shown in Fig. 2. Specifically, we design our 3D branch to take a 3D point cloud as input and to predict their 3D features. The semantically meaningful mask embeddings produced from our 2D branch are used as linear classifiers to assign class probability to these 3D features. Their corresponding 2D masks are lifted to 3D based on pixel-point correspondence and used to force the consistency learning of the 2D and 3D branch.

We evaluate *Diff2Scene* quantitatively on ScanNet [17], ScanNet200 [69], Matterport 3D [7] and Replica [77] for open-vocabulary 3D semantic segmentation and qualitatively on Nr3D [1] for visual grounding tasks. Our experimental results show that *Diff2Scene* outperforms state-of-the-art models [62] on all the four semantic segmentation datasets and achieves promising results on visual grounding tasks. In summary, we make the following contributions:

- To the best of our knowledge, we are the first to leverage text-image diffusion to perform open-vocabulary 3D semantic segmentation.
- We propose a novel mask distillation method to train a 3D mask prediction model by distilling knowledge from the Mask2Former style 2D segmentation model.
- The proposed method achieves state-of-the-art performance on several open-vocabulary 3D semantic segmentation and visual grounding benchmarks.

2 Related Work

Closed-vocabulary 3D semantic segmentation. In 3D semantic segmentation, a semantic category is assigned for each 3D point. It has been long studied [3–5, 20, 21, 26, 31, 32, 42, 45, 49, 57, 63, 64, 82, 86, 95] due to its importance in computer vision and robotics applications. One challenge of this task is that 3D point clouds are not in a regular structured format; network architectures that work well for 2D tasks cannot handle 3D point clouds effectively. As a result, most of the early studies focus on designing effective and efficient network architectures that are suitable for 3D point clouds [16, 20, 21, 26, 31, 32, 45, 63, 64, 83]. This line of work achieved great success and significantly improves the results of

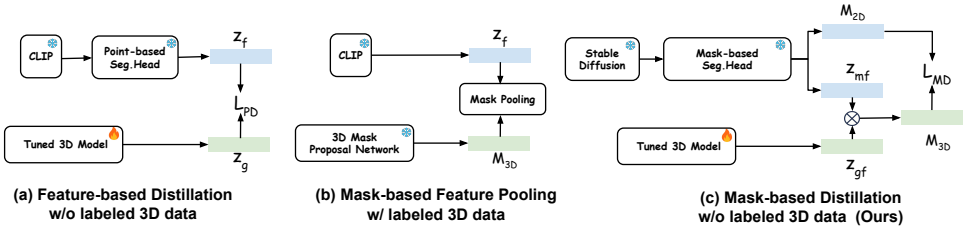

Fig. 2: Illustration of open-vocabulary 3D perception methods. L_{PD} and L_{MD} denote point-based distillation loss and mask-based distillation loss. M_{3D} denote a set of predicted 3D masks; M_{2D} and Z_{mf} denote a set of predicted 2D masks and their semantic embeddings; Z_{gf} denote the high-resolution 3D feature map. (a) Directly minimizing the per-point feature distance between the CLIP-based model and the tuned 3D model [62]. (b) Directly using a 3D mask proposal network trained on labeled 3D data to produce class-agnostic masks, and then pool corresponding representations from the CLIP feature map [79]. (c) The proposed mask distillation approach, namely *Diff2Scene*, that uses Stable Diffusion and performs mask-based distillation. *Diff2Scene* leverages the semantically-rich mask embeddings from 2D foundation models and geometrically accurate masks from the tuned 3D model, and thus achieves superior performance compared to previous methods.

3D semantic segmentation. Another challenge is the lack of large scale data with ground truth annotations. Due to the intensive labeling effort and high cost of data annotation [69], the available datasets for 3D semantic segmentation are usually small in scale. In the absence of large scale data, early studies usually target solving this problem in a closed-vocabulary setting, where the trained model can only predict categories that appear during training. To mitigate the scale limitation of existing datasets, a handful of works [12–15, 52, 60, 91] have applied zero-shot learning in 3D scene understanding tasks. [12–15, 91] focused on 3D point classification task and [52, 60] tried to address the 3D semantic segmentation problem. However, these zero-short methods still require ground truth annotations for a certain amount of 3D point clouds.

Open-vocabulary 3D segmentation. The recent progress of large-scale vision and language representation learning [2, 28, 39, 48, 65, 90] has advanced the study of semantic and instance segmentation in an open-vocabulary setting. [25, 47, 51] first explored open-vocabulary 2D semantic segmentation. They proposed aligning per-pixel features [47] or features from mask regions [25, 51] with the corresponding text embedding. Following these works, [8, 19, 24, 33, 37, 59, 62, 74, 75] focus on 3D semantic segmentation in an open-vocabulary setting. Among them, [8, 24, 33, 37, 59, 71, 74] project 3D points to 2D images and solve the 3D problem in the 2D space, instead of targeting the 3D open-vocabulary semantic segmentation directly. As [19] pointed out, the projection from 3D to 2D has information loss and the solution is suboptimal.

To make better use of information from the 3D point cloud, [19], [18] and [62] proposed to directly applying semantic segmentation on the 3D point cloud. [19] and its extension [18] proposed associating captions generated for 2D images to corresponding 3D point clouds to build the pseudo-ground truth captions for 3D point clouds. A neural network is trained to associate the 3D point cloud with these pseudo labels through contrastive loss. Similar to the zero-shot setting, [19] evaluated their model in a leave-one-out fashion, which still requires annotations for 3D point cloud. Inspired by the strong open-vocabulary ability of large-scale vision and language models, Peng *et al*. [62] proposed distilling knowledge to a 3D point cloud model. They trained 3D semantic segmentation model by only distilling the knowledge from a CLIP-style [65] 2D open-vocabulary semantic segmentation model [25, 47]. They demonstrated that without training with any ground truth labels, the model can achieve great performance on many open-vocabulary tasks. However, we observe that [62] is strongly limited by the 2D open-vocabulary semantic segmentation models used as the teacher. Its performance on rare classes that are not used in training these models are not satisfactory. Our method follows this idea by distilling the knowledge of 2D open-vocabulary semantic segmentation model to a 3D model. In constrast to the approach in [62], we use a diffusion-based 2D open vocabulary semantic segmentation model [87] as the teacher model.

Diffusion models for scene understanding. The last few years have witnessed the success of diffusion models in image generation [68, 70]. Recent studies also observed the diffusion models are strong representation learners [43, 60, 61, 89]. As a result, researchers have applied it to many understanding tasks such as image classification [46], object detection [9], image semantic segmentation [6, 38, 87], instance segmentation [50], human pose estimation [23, 29, 76], action segmentation [53], camera pose estimation [84], to name a few, and achieved great success. Especially, [87] and [50] showed that Stable Diffusion [68], whose internal representation being well correlated with text embedding, has strong open-vocabulary abilities for understanding tasks. Inspired by this, we are the first to apply text-to-image diffusion models to open vocabulary 3D semantic segmentation task.

3 Diff2Scene

We introduce *Diff2Scene*, an open-vocabulary 3D semantic understanding method. Similar to [62], our proposed model operates in a zero-shot fashion, where no ground truth 3D annotations are needed during training.

3.1 Overview

An overview of *Diff2Scene* is shown in Fig. 3. It takes posed RGB images and the reconstructed 3D point cloud as model inputs. The model predicts the semantic label for each 3D point. *Diff2Scene* has two branches. The 2D branch is designed

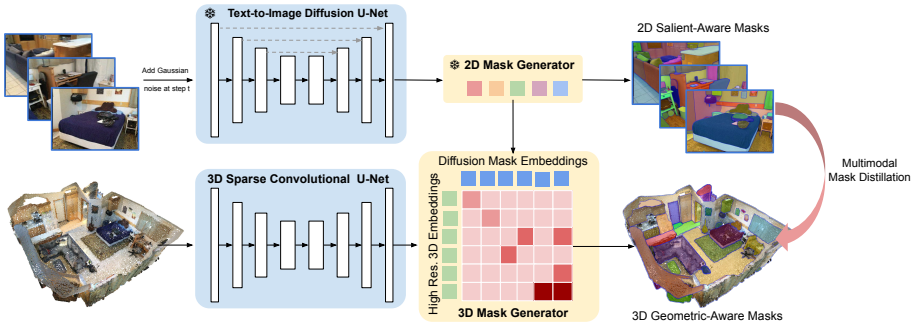

Fig. 3: Overview of our method. We propose *Diff2Scene*, an open-vocabulary 3D semantic understanding model. *Diff2Scene* contains two branches. The 2D branch is designed to be a diffusion-based 2D semantic segmentation model. It accepts a 2D image as input and predicts a set of 2D probabilistic masks with corresponding semantically-rich mask embeddings. The 3D branch utilizes the point cloud and 2D mask embeddings as input. The 2D mask embeddings are used as "semantic queries" to generate corresponding 3D probabilistic masks. The model learns salient patterns from the RGB images and geometric information from the point clouds.

to be an open-vocabulary 2D semantic segmentation model. It leverages text-to-image generative model [68] which is pre-trained on massive text-image pairs. The model takes a 2D image as input to predict a set of 2D probabilistic masks with their corresponding 2D mask embeddings. Thanks to the generative pre-training process with large-scale text-image pairs, the 2D mask embeddings are semantically rich. The model leverages the salient patterns in RGB images to produce the 2D salient masks. The 3D branch takes the point cloud and the 2D mask embeddings as inputs. The 2D mask embeddings are used as linear classifiers to assign class probabilities to each of 3D features output from the 3D branch, resulting in a 3D probabilistic mask termed as geometric masks. To predict the per-point semantic class, the model first computes the per-mask category logits for both salient and geometric masks. Then we ensemble the per-mask logits for those two types of masks. In the way, the model can learn salient patterns from the RGB images and geometric information from the point clouds.

3.2 2D Semantic Understanding Model

One challenge of 3D semantic understanding is the severe scarcity of 3D point clouds with groundtruth labels. To tackle the challenge brought by limited training data, vision-language foundation models have been used to transfer semantically-rich 2D features into the 3D space [37, 62, 79]. [79] used on a model trained on labeled 3D data to produce class-agnostic masks, and then pooled the corresponding 2D representations as the mask embeddings. On the other hand, [62] proposed to leverage a pre-trained 2D semantic segmentation model as feature extractor to perform open-vocabulary 3D segmentation, and no ground

truth 3D annotations are needed during training. In this work, we follow the setting in [62] to reduce the 3D annotation efforts.

The 2D segmentation model consists of an image backbone ϕ which is a foundation model pretrained on large-scale text-image pairs; and a segmentation head σ to predict the semantic embedding. There are multiple design options for the 2D backbone ϕ and segmentation head σ. (1) The 2D backbone could either be **contrastively pretrained** or **generatively pre-trianed**. The popular frameworks for contrastive representation learning include CLIP [65] and ALIGN [40]. On the other hand, a few works [85,87,92] have demonstrated promising performance by using generatively pre-trained representations for perception task. The feature representations from Diffusion U-Net blocks are extracted for different downstream tasks.

Once feature representations from text-image foundation models are extracted, a segmentation head is added upon those features to predict the per-point semantic classes. The segmentation problem could be formulated as **pixel-based classification** or **mask-based classification**. For pixel-based classification [25,47], the intermediate output of segmentation head is of shape $H \times W \times C$, where H and W is image height and width, and C is the dimension of feature embedding. For mask-based classification [10, 11, 87], the segmentation head takes the 2D feature map $\mathbf{F}^{2d}$ and N fixed mask queries $\{q_i\}_{i=1}^{N}$ as input. The intermediate output is N 2D probabilistic masks $\{\mathcal{B}_i^{2d}\}_{i=1}^{N}$ and their corresponding mask embeddings $\{f_i^{2d}\}_{i=1}^{N}$.

In this work, we choose diffusion model as the feature backbone ϕ, considering its strong localization ability brought by generative pre-training. Besides, we leverage mask-based segmentation head for its intrinsic nature that decouples mask and its semantic representations. This is intuitively suitable for leveraging semantically-rich embeddings from 2D foundation models, and further learn geometrically-accurate masks from the 3D branch.

3.3 Geometry-Aware 3D Mask Model

While mask-based segmentation has achieved promising performance in fully-supervised setting [10,11,73], it has been rarely explored to transfer the learned mask-level representations into another domain. On the other hand, the point-based feature representations from 2D foundation model can be naively distilled by minimizing the per-point feature distance. For example, [62] proposed to train a 3D model to predicts 3D semantic meaningful features by distilling pixel aligned 2D features. However, similar methods are not applicable in our proposed method. First of all, our 2D semantic understanding model uses a mask-based segmentation head which does not provide semantically-rich features in the pixel level. Secondly, the backbone of our 2D semantic understanding model is a frozen stable diffusion model [68] which is designed to generate realistic images with rich details and not tuned for semantic segmentation tasks. The per-pixel features extracted from it are not feasible to supervise the training of our 3D mask

model†. In the following, we introduce our proposed mask distillation which is tailored to distill knowledge from the mask-based 2D foundation model to the geometry-aware 3D mask model.

The mask-based 2D foundation model predicts N 2D probabilistic masks $\{\mathcal{B}_i^{2d}\}_{i=1}^N$ and their corresponding mask embeddings $\{f_i^{2d}\}_{i=1}^N$. Specifically, $\mathcal{B}_i^{2d}$ represents a probabilistic map whose elements represent the probability of the corresponding pixel being foreground. We first compute the pixel-point correspondence following [62]. Subsequently, a set of 3D probabilistic masks $\{\mathcal{B}_i^{3d}\}_{i=1}^N$ can be generated by lifting the 2D masks $\{\mathcal{B}_i^{2d}\}_{i=1}^N$ to 3D space based on the pixel-point correspondence. We proposed a novel mask distillation which distills information from both 3D probabilistic masks $\{\mathcal{B}_i^{3d}\}_{i=1}^N$ and the corresponding semantic rich mask embeddings $\{f_i^{2d}\}_{i=1}^N$ generated from the 2D branch. Specifically, we train a Minkowski network [16] as the 3D mask prediction model to generate geometry-aware 3D masks. The 3D point cloud is quantized into voxels by averaging the pixels within each voxel to save memory and reduce computes. The 3D mask prediction model generates a 3D feature to represent each voxel and this feature is assigned to all points within the voxel. This produces a full feature map $\mathbf{F}^{3d} \in \mathbb{R}^{M \times D}$ for the point cloud, where D is the dimension of the 3D feature. The semantic rich 2D mask embeddings $\{f_i^{2\mathrm{d}}\}_{i=1}^N$ are used as linear classifiers to compute the logits $\mathcal{S}_i \in \mathbb{R}^M$ of a 3D feature belonging to the corresponding class:

$$\mathcal{S}_i = \langle \mathbf{F}^{3d}, f_i^{2d} \rangle, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes inner product. The 3D probabilistic mask $\mathcal{B'}_i^{3d}$ is then generated by applying the sigmoid function on $\mathcal{S}_i$. We propose a multimodal mask distillation loss to train our 3D mask generator:

$$\mathcal{L} = \sum_{i=1}^{N} 1 - \cos(\mathcal{B'}_i^{3d}, \mathcal{B}_i^{3d}). \tag{2}$$

The distillation loss aims at forcing the 2D and 3D branch to make consistent predictions. It serves as an implicit distillation objective to make the 3D model learn high-resolution, semantically-rich feature representations.

3.4 Open-Vocabulary Inference

During inference, *Diff2Scene* takes a 3D point cloud and its multiview 2D images as inputs. The 2D semantic understanding model consumes the 2D images and generates a set of 2D probabilistic masks $\{\mathcal{B}_i^{2d}\}_{i=1}^N$ with their corresponding mask embeddings $\{f_i^{2\mathrm{d}}\}_{i=1}^N$, where $f_i^{2d} \in \mathbb{R}^D$. The 3D mask model takes the 3D point cloud and the mask embeddings $\{f_i^{2\mathrm{d}}\}_{i=1}^N$ as inputs to predict the 3D probabilistic mask $\{\mathcal{B'}_i^{3d}\}_{i=1}^N$. To ground a semantic label c to the 3D point cloud, we first apply the same idea from [87] to compute the geometric mean (denoted as p_i^c) of label probabilities from diffusion and discriminative models for each 2D

†The 3D mask model trained to distill these features does not converge.

mask $\{\mathcal{B}_i^{2d}\}_{i=1}^N$. Next, the label probabilities $\mathbf{p}^c$ are assigned to 3D points via the following equation:

$$\mathbf{p}^c = \lambda \sum_{i=1}^{N} p_i^c * \mathcal{B}_i^{3d} + (1-\lambda) \sum_{i=1}^{N} p_i^c * \mathcal{B'}_i^{3d}, \tag{3}$$

where $\lambda = 0.5$. When multiple labels can be assigned to a 3D point, the label with the highest probability from Eq. 3 is taken.

4 Experiment

We conduct a series of experiments to demonstrate the effectiveness of *Diff2Scene* on a variety of zero-shot 3D scene understanding benchmarks. We first evaluate the proposed model on zero-shot open-vocabulary semantic segmentation tasks following the evaluation protocol of [62]. We then perform comprehensive ablation studies to validate our designs. Finally, we qualitatively demonstrate the strong ability of the proposed model for open-vocabulary 3D segmentation and grounding complicated compositional text queries.

4.1 Datasets

We use ScanNet [17], Matterport3D [7], ScanNet200 [69] and Replica [77] for the open-vocabulary 3D semantic segmentation task. We provide qualitative analysis of the visual grounding task on Nr3D [1]. Except for Replica, point clouds and multi-view images in the training split without ground truth annotations are used for model training. As Replica does not provide the training data, we perform training on ScanNet and perform evaluation on Replica, following the setting in [79].

ScanNet is one of the largest 3D semantic segmentation dataset. It provides 80,554 images from 1201 scans for training and 21,300 images from 312 scans for testing with 20 semantic labels.

Matterport3D is a large scale RGB-D dataset containing 10,800 panoramic views from 194,000 RGB-D images of 90 building-scale scenes. It splits 61 scenes for training, 11 scenes for validation and 18 for testing. We train our 3D branch using the images in the training splits and report the results on test split.

ScanNet200 has 200 semantic labels with long-tailed classes. It also provides a grouping of the 200 categories based on the number of labeled surface points in the training set, resulting in 3 subsets: head, common, and tail. This enables us to evaluate the performance of our method on the long-tail distribution, making ScanNet200 a natural choice as an evaluation dataset. We report the mean intersection over union (mIoU) metric on the validation set consisting of 312 scenes following the split in [62, 69, 78].

Table 1: Comparison to state-of-the-art models. We report mIoU for all benchmarks. Best results in zero-shot, open-vocabulary setting are shown in bold.

	ScanNet	**Matterport3D**	**ScanNet200**				**Replica**		
	All	All	Head	Common	Tail	All	Head	Tail	All
Fully-supervised									
TangentConv [81]	40.9	-	-	-	-	-	-	-	-
TextureNet [34]	54.8	-	-	-	-	-	-	-	-
SFSS-MMSI [16]	-	35.9	-	-	-	-	-	-	-
CSC-Pretrain [30]	-	-	45.5	17.1	7.9	24.9	-	-	-
SupCon [93]	-	-	48.6	19.2	10.3	26.0	-	-	-
LGround [69]	-	-	48.5	18.4	10.6	27.2	-	-	-
MinkowskiNet [16]	69.0	54.2	46.3	15.4	10.2	25.3	-	-	-
Zero-shot, open-vocabulary									
MSeg Voting [44]	31.0	33.4	-	-	-	-	-	-	-
ConceptFusion [37]	33.3	-	17.5	6.3	2.8	8.8	11.6	3.5	4.6
OpenMask3D [79]	34.0	-	19.6	7.5	4.5	10.5	13.2	3.4	4.8
OpenScene (2D) [62]	41.4	32.4	21.9	10.8	5.5	12.7	33.4	11.5	14.5
OpenScene (3D) [62]	46.0	41.3	17.6	0.0	0.0	6.3	32.6	7.7	11.1
OpenScene (2D/3D) [62]	47.5	42.6	20.0	9.7	5.1	11.6	34.2	11.9	14.9
Diff2Scene (Ours)	**48.6**	**45.5**	**25.6**	**11.5**	**6.9**	**14.2**	**46.2**	**12.9**	**17.5**

Replica contains 51 categories, and we further split those categories into head and tail sets based on their appearance frequency. We report the mIoU on the *office0*, *office1*, *office2*, *office3*, *office4*, *room0*, *room1*, and *room2*.

Nr3D is a 3D visual grounding dataset which contains diverse text prompts. To further evaluate the ability of our model to distinguish between objects in the same class but with different attributes, we perform qualitative evaluation on the visual grounding dataset Nr3D [1]. We perform zero-shot evaluation on the validation set without training on any labeled data for the visual grounding task.

4.2 Baseline Methods

We compare *Diff2Scene* with the current state-of-the-art fully-supervised 3D semantic segmentation models including TangentConv [81], TextureNet [34], SFSS-MMSI [16], CSC-Pretrain [30], SupCon [93], LGround [69] and MinkowskiNet [16] on the 3D semantic segmentation benchmark. We also compare our model against OpenScene [62] and ConceptFusion [37], the recently proposed open-vocabulary 3D semantic understanding model. For OpenScene [62], we compare with its OpenSeg [25] variant which has the same feature and pre-trained datasets for a fair comparison. We also compare our model with its three different variants (2D Fusion, 3D Distill, and 2D/3D Ensemble). Besides, we adapt the state-of-the-art 3D instance segmentation model OpenMask3D [79] for comparison on the 3D semantic segmentation benchmark.

4.3 Implementation Details

We use posed multi-view RGB images and 3D point clouds for all the datasets. ODISE [87], which consists of a diffusion backbone and mask-based segmentation head, is used as the model in our 2D branch. It uses a stable diffusion model [68] pre-trained on Laion-5B [72] as the feature backbone. The dimensions for diffusion and CLIP features are 256 and 768 respectively. The number of queries of Mask2Former [10] is 100. Similar to OpenScene [62], we use MinkowskiNet18A [16] as the model in our 3D branch to extract 3D features from the 3D point clouds. Our 3D model is trained for 200 epochs with a batch size of 8. Adam optimizer [41] is used with a learning rate of 0.0001 and polynomial learning rate policy is used as the learning rate scheduler with power 0.9. During inference, text-embeddings are computed by the ViT-L/14 CLIP model [65] for each of the semantic categories and grounding queries. We use the same pre-processing step and pre-trained dataset as OpenScene [62] (OpenSeg [25]) for a fair comparison.

4.4 Quantitative Results

Evaluation on zero-shot 3D semantic segmentation. We first compare our method with the state-of-the-art open-vocabulary scene understanding models and fully-supervised 3D segmentation models. We report the mIoU for Scannet, Matterport3D, Scannet200, and Replica in Table 1. We find that our method achieves better results than the state-of-the-art open-vocabulary models and their variants on all the benchmarks. Besides, although our zero-shot model has noticeable performance drop compared with fully-supervised model, the gap of tail categories between the proposed method and those methods are relatively small (*e.g.* 6.9 *vs.* 7.9 from CSC-Pretrain) on Scannet200. This demonstrates the strong potential of the proposed method for long-trailed 3D semantic segmentation tasks.

Generalization to unseen dataset. To test the generalization ability of our proposed model, we evaluate it on an unseen dataset Replica [77] and report the results in Table 1. The results shown that our proposed method significantly outperforms the state-of-the-art models on head, tail and all categories in Replica. This demonstrates the strong generalization ability of the proposed model on novel datasets.

Effectiveness of Different Distillation Settings. We compare our mask-based distillation method with point-based ones under different settings and report the performance of the 3D branch on Replica [77] in Table 2. The supervisions for point-based method include: (1) Fine-tuned CLIP feature, which follows the same setting as OpenScene [62]; (2) Frozen diffusion feature extracted from the last layer of diffusion U-Net block. We observe that distilling frozen diffusion features does not converge. Our proposed method, by introducing the semantic meaningful mask embedding output from the 2D branch as a fixed classifier, significantly boost the performance of the 3D branch.

Ablation studies. We conduct ablation studies using the Replica dataset [77] and show the results in Table 3. We first analyze the effectiveness of combining

Table 2: Effectiveness of Different Distillation Settings. We report mIoU of different methods on the Replica [77] dataset.

Setting	Distillation Type	Head	Tail	All
fine-tuned CLIP feature [62]	Point-based	32.6	7.7	11.1
frozen diffusion feature	Point-based	Divergence		
multimodal mask distillation (ours)	Mask-based	**43.3**	**8.0**	**12.8**

Table 3: Performance of different model ablations. We observe that each component of our model gains consistent improvements.

Method	mIoU
Our full model	17.5
Without 2D (salient) mask	12.8
Without 3D (geometric) mask	16.5
Without discriminative (CLIP) features	15.5
Without generative (Stable Diffusion) features	15.3

2D and 3D masks using equation 3. We observe that compared with using salient or geometric mask only, using both types of masks achieves the best performance. This is intuitive as both salient patterns and geometric information are helpful to segment accurate class boundaries. We then analyze the effectiveness of different semantic features. We find that discriminative and diffusion features serve as strong complementary to each other. We also observe that using those two types of semantic features jointly can significantly outperform using any of them alone.

4.5 Qualitative analysis

Visualizations of zero-shot semantic segmentation. In Fig. 4, we provide qualitative analysis of our approach and OpenScene for the zero-shot 3D semantic segmentation task. Compared with OpenScene, our model generates coherent and consistent masks (*e.g.*, the table mask in first column and the bed mask in third column) thanks to the mask-instance representations. It predicts accurate semantic labels for both head and tail categories by leveraging both CLIP and diffusion features.

Visualizations of visual grounding results. We provide qualitative analysis of our approach and OpenScene for the zero-shot visual grounding task in Fig. 5. We observe that our model can accurately identify objects given complicated text queries. It demonstrates that the proposed method, *Diff2Scene*, has good capability at the following types of queries. Fig. 5 (a) describes object shape and color, and even in comparative degree (*It's the shorter, red box*); Fig. 5 (b) describes a rare object (*rack*) and its surrounded object with surface appearance descriptions (*wrinkled towel*); Fig. 5 (c) describes the relative location of the

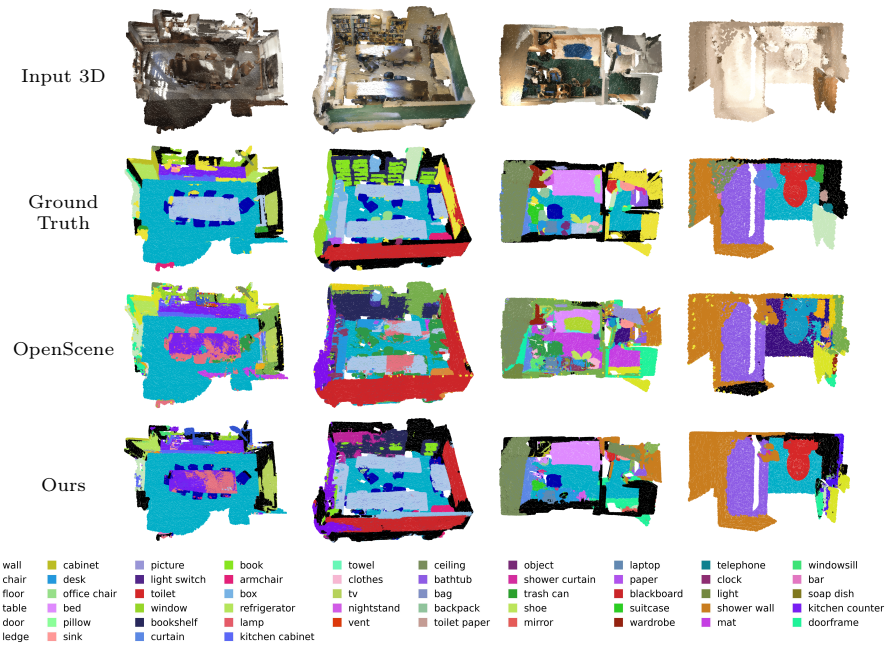

Fig. 4: Qualitative results from our model and OpenScene on zero-shot semantic segmentation. We visualize the segmentation results on the validation set of ScanNet200 [69]. We observe that our model can predict coherent masks with accurate semantic labels compared to OpenScene for both head and tail categories.

object (*next to the desk*); Fig. 5 (d) describes the usage of the object (*recycling*). In addition, we can see that given vague usage descriptions without common category names like *trash bin* in the text prompts, the model can still accurately identify the object.

5 Conclusion

In this paper, we investigate the problem of leveraging frozen representations from large text-to-image diffusion models for open-vocabulary 3D semantic understanding. *Diff2Scene* sets a new state-of-the-art in the zero-shot 3D semantic segmentation task and shows promising performance in the visual grounding task. Our method also shows outstanding generalization ability towards unseen datasets and novel text queries. It provides a new way to effectively leverage generative text-to-image foundation models for 3D semantic scene understanding tasks.

There are several limitations of the proposed model. First, while our model achieves better performance compared to existing methods in small objects, it still misclassified some small and rare categories (*e.g.* rail). Second, we observe that the model can be easily confused by fine-grained categories that with similar

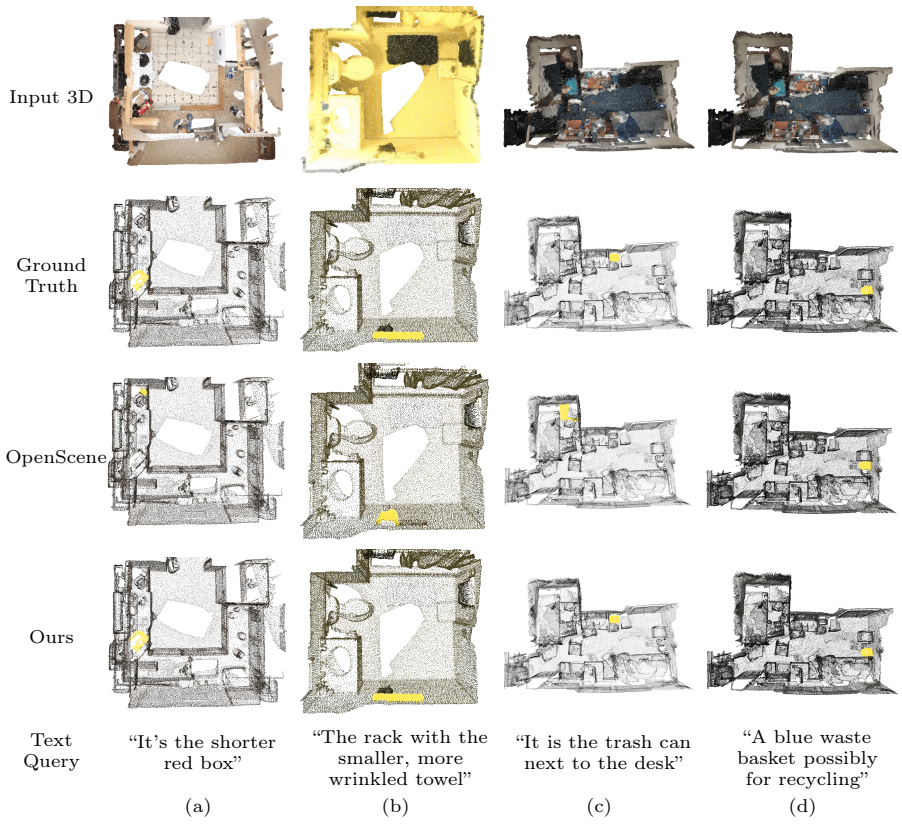

Fig. 5: Qualitative results from our model and OpenScene on zero-shot visual grounding. Our open-vocabulary semantic understanding model is capable of handling different types of novel and compositional queries. Novel object classes as well as objects described by colors, shapes, appearances, locations, and usages are successfully retrieved by our method. Note that the located points are colored in yellow.

semantic meaning. For example, the model sometimes wrongly classifies points of windowsill to the window class. In future work, it will be interesting to design models that can accurately distinguish between fine-grained categories in the open-vocabulary setting.

Acknowledgements

We sincerely thank Amir Hertz, Weicheng Kuo, and Lijun Yu for the helpful discussions.

References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.J.: ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In: ECCV (2020) 3, 9, 10
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: NeurIPS (2022) 2, 4
3. Anand, A., Koppula, H.S., Joachims, T., Saxena, A.: Contextually guided semantic labeling and search for 3d point clouds. In: IJRR (2011) 3
4. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: CVPR (2016) 3
5. Atzmon, M., Maron, H., Lipman, Y.: Point convolutional neural networks by extension operators. ACM Transactions on Graphics (2018) 3
6. Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. In: ICLR (2022) 5
7. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. In: 3DV (2017) 3, 9
8. Chen, B., Xia, F., Ichter, B., Rao, K., Gopalakrishnan, K., Ryoo, M.S., Stone, A., Kappler, D.: Open-vocabulary queryable scene representations for real world planning. arXiv preprint arXiv:2209.09874 (2022) 4
9. Chen, S., Sun, P., Song, Y., Luo, P.: Diffusiondet: Diffusion model for object detection. In: ICCV (2023) 5
10. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR (2022) 2, 3, 7, 11
11. Cheng, B., Schwing, A.G., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. In: NeurIPS (2021) 7
12. Cheraghian, A., Rahman, S., Campbell, D., Petersson, L.: Transductive zero-shot learning for 3d point cloud classification. In: WACV (2020) 4
13. Cheraghian, A., Rahman, S., Campbell, D., Petersson, L.: Mitigating the hubness problem for zero-shot learning of 3d objects. In: BMVC (2019) 4
14. Cheraghian, A., Rahman, S., Chowdhury, T.F., Campbell, D., Petersson, L.: Zero-shot learning on 3d point cloud objects and beyond. IJCV (2022) 4
15. Cheraghian, A., Rahman, S., Petersson, L.: Zero-shot learning of 3d point cloud objects. In: MVA (2019) 4
16. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: CVPR (2019) 2, 3, 8, 10, 11
17. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017) 3, 9
18. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: Lowis3d: Language-driven open-world instance-level 3d scene understanding. arXiv preprint arXiv:2308.00353 (2023) 5
19. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: Pla: Language-driven open-vocabulary 3d scene understanding. In: CVPR (2023) 4, 5
20. Engelmann, F., Kontogianni, T., Hermans, A., Leibe, B.: Exploring spatial context for 3d semantic segmentation of point clouds. In: ICCV workshop (2017) 3

21. Engelmann, F., Kontogianni, T., Schult, J., Leibe, B.: Know what your neighbors do: 3d semantic segmentation of point clouds. In: ECCV workshop (2019) 3
22. Fan, J., Zheng, P., Li, S.: Vision-based holistic scene understanding towards proactive human–robot collaboration. Robotics and Computer-Integrated Manufacturing **75**, 102304 (2022) 2
23. Feng, R., Gao, Y., Tse, T.H.E., Ma, X., Chang, H.J.: Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In: ICCV (2023) 5
24. Gadre, S.Y., Wortsman, M., Ilharco, G., Schmidt, L., Song, S.: Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In: CVPR (2023) 4
25. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: ECCV (2022) 2, 4, 5, 7, 10, 11
26. Graham, B., Engelcke, M., Maaten, L.v.d.: 3d semantic segmentation with submanifold sparse convolutional networks. In: CVPR (2018) 3
27. Han, L., Zheng, T., Zhu, Y., Xu, L., Fang, L.: Live semantic 3d perception for immersive augmented reality. IEEE Trans. visualization and computer graphics **26**(5), 2012–2022 (2020) 2
28. He, Q., Peng, J., Jiang, Z., Wu, K., Ji, X., Zhang, J., Wang, Y., Wang, C., Chen, M., Wu, Y.: Unim-ov3d: Uni-modality open-vocabulary 3d scene understanding with fine-grained feature representation. IJCAI (2024) 4
29. Holmquist, K., Wandt, B.: Diffpose: Multi-hypothesis human pose estimation using diffusion models. In: ICCV (2023) 5
30. Hou, J., Graham, B., Nießner, M., Xie, S.: Exploring data-efficient 3d scene understanding with contrastive scene contexts. In: CVPR (2021) 10
31. Hu, Z., Bai, X., Shang, J., Zhang, R., Dong, J., Wang, X., Sun, G., Fu, H., Tai, C.L.: Vmnet: Voxel-mesh network for geodesic-aware 3d semantic segmentation. In: ICCV (2021) 3
32. Hua, B.S., Tran, M.K., Yeung, S.K.: Pointwise convolutional neural networks. In: CVPR (2018) 3
33. Huang, C., Mees, O., Zeng, A., Burgard, W.: Visual language maps for robot navigation. In: ICRA (2023) 4
34. Huang, J., Zhang, H., Yi, L., Funkhouser, T., Nießner, M., Guibas, L.J.: Texturenet: Consistent local parametrizations for learning from high-resolution signals on meshes. In: CVPR (2019) 10
35. Huang, S., Chen, Y., Jia, J., Wang, L.: Multi-view transformer for 3d visual grounding. In: CVPR (2022) 2
36. Huang, Z., Lv, C., Xing, Y., Wu, J.: Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding. IEEE Sensors Journal **21**(10), 11781–11790 (2020) 2
37. Jatavallabhula, K., Kuwajerwala, A., Gu, Q., Omama, M., Chen, T., Li, S., Iyer, G., Saryazdi, S., Keetha, N., Tewari, A., Tenenbaum, J., de Melo, C., Krishna, M., Paull, L., Shkurti, F., Torralba, A.: Conceptfusion: Open-set multimodal 3d mapping. In: Robotics science and systems (2023) 2, 4, 6, 10
38. Ji, Y., Chen, Z., Xie, E., Hong, L., Liu, X., Liu, Z., Lu, T., Li, Z., Luo, P.: Ddp: Diffusion model for dense visual prediction. In: ICCV (2023) 5
39. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021) 2, 4
40. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021) 7

41. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 11
42. Koppula, H., Anand, A., Joachims, T., Saxena, A.: Semantic labeling of 3d point clouds for indoor scenes. In: NeurIPS (2011) 3
43. Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. In: ICLR (2023) 2, 5
44. Lambert, J., Liu, Z., Sener, O., Hays, J., Koltun, V.: MSeg: A composite dataset for multi-domain semantic segmentation. In: CVPR (2020) 10
45. Landrieu, L., Simonovsky, M.: Large-scale point cloud semantic segmentation with superpoint graphs. In: CVPR (2018) 3
46. Li, A.C., Prabhudesai, M., Duggal, S., Brown, E., Pathak, D.: Your diffusion model is secretly a zero-shot classifier. In: ICCV (2023) 5
47. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: ICLR (2022) 2, 4, 5, 7
48. Li, J., Selvaraju, R.R., Gotmare, A.D., Joty, S.R., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: NeurIPS (2021) 2, 4
49. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B.: Pointcnn: Convolution on x-transformed points. In: NeurIPS (2018) 3
50. Li, Z., Zhou, Q., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Open-vocabulary object segmentation with diffusion models. In: ICCV (2023) 5
51. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: CVPR (2023) 4
52. Liu, B., Deng, S., Dong, Q., Hu, Z.: Language-level semantics conditioned 3d point cloud segmentation. arXiv prepreint arXiv:2107.00430 (2022) 4
53. Liu, D., Li, Q., Dinh, A.D., Jiang, T., Shah, M., Xu, C.: Diffusion action segmentation. In: ICCV (2023) 5
54. Liu, K., Zhan, F., Zhang, J., Xu, M., Yu, Y., Saddik, A.E., Theobalt, C., Xing, E., Lu, S.: Weakly supervised 3d open-vocabulary segmentation. In: NeurIPS (2023) 3
55. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: ECCV (2022) 2
56. Liu, Z., Qi, X., Fu, C.W.: 3d-to-2d distillation for indoor scene parsing. In: CVPR (2021) 3
57. Lu, Y., Rasmussen, C.: Simplified markov random fields for efficient semantic labeling of 3d point clouds. In: ICIRS (2012) 3
58. Ma, Z., Hong, J., Gul, M.O., Gandhi, M., Gao, I., Krishna, R.: Crepe: Can vision-language foundation models reason compositionally? arXiv preprint arXiv:2212.07796 (2023) 2
59. Mazur, K., Sucar, E., Davison, A.: Feature-realistic neural fusion for real-time, open set scene understanding. In: ICRA (2023) 4
60. Michele, B., Boulch, A., Puy, G., Bucher, M., Marlet, R.: Generative zero-shot learning for semantic segmentation of 3D point cloud. In: 3DV (2021) 2, 4, 5
61. Mittal, S., Abstreiter, K., Bauer, S., Schölkopf, B., Mehrjou, A.: Diffusion based representation learning. In: ICML (2023) 2, 5
62. Peng, S., Genova, K., Jiang, C.M., Tagliasacchi, A., Pollefeys, M., Funkhouser, T.: Openscene: 3d scene understanding with open vocabularies. In: CVPR (2023) 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
63. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR (2017) 3

64. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: NeurIPS (2017) 3
65. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021) 2, 4, 5, 7, 11
66. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 2
67. Roh, J., Desingh, K., Farhadi, A., Fox, D.: Languagerefer: Spatial-language model for 3d visual grounding. In: Conference on Robot Learning. pp. 1046–1056. PMLR (2022) 2
68. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 2, 5, 6, 7, 11
69. Rozenberszki, D., Litany, O., Dai, A.: Language-grounded indoor 3d semantic segmentation in the wild. In: ECCV (2022) 3, 4, 9, 10, 13
70. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Gontijo-Lopes, R., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022) 5
71. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022) 4
72. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022) 2, 11
73. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In: ICRA (2023) 2, 7
74. Shafiullah, N.M.M., Paxton, C., Pinto, L., Chintala, S., Szlam, A.: CLIP-fields: Weakly supervised semantic fields for robotic memory. In: CoRL Workshop on Language and Robotics (2022) 4
75. Shah, D., Osinski, B., Ichter, B., Levine, S.: LM-nav: Robotic navigation with large pre-trained models of language, vision, and action. In: CoRL (2022) 4
76. Shan, W., Liu, Z., Zhang, X., Wang, Z., Han, K., Wang, S., Ma, S., Gao, W.: Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In: ICCV (2023) 5
77. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019) 3, 9, 11, 12
78. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In: NeurIPS (2023) 2, 9
79. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: Openmask3d: Open-vocabulary 3d instance segmentation. arXiv preprint arXiv:2306.13631 (2023) 2, 3, 4, 6, 9, 10
80. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. arXiv preprint arXiv:2306.03881 (2023) 2
81. Tatarchenko, M., Park, J., Koltun, V., Zhou., Q.Y.: Tangent convolutions for dense prediction in 3D. CVPR (2018) 10

82. Tchapmi, L.P., Choy, C.B., Armeni, I., Gwak, J., Savarese, S.: Segcloud: Semantic segmentation of 3d point clouds. In: 3DV (2017) 3
83. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: ICCV (2019) 3
84. Wang, J., Rupprecht, C., Novotny, D.: Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In: ICCV (2023) 5
85. Wang, J., Li, X., Zhang, J., Xu, Q., Zhou, Q., Yu, Q., Sheng, L., Xu, D.: Diffusion model is secretly a training-free open vocabulary semantic segmenter. arXiv preprint arXiv:2309.02773 (2023) 7
86. Wang, T., Li, J., An, X.: An efficient scene semantic labeling approach for 3d point cloud. In: ITSC (2015) 3
87. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: CVPR (2023) 2, 3, 5, 7, 8, 11
88. Xu, Z., He, Z., Wu, J., Song, S.: Learning 3d dynamic scene representations for robot manipulation. arXiv preprint arXiv:2011.01968 (2020) 2
89. Yang, X., Wang, X.: Diffusion model as representation learner. In: ICCV (2023) 2, 5
90. Zhang, J., Dong, R., Ma, K.: Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. arXiv preprint arXiv:2303.04748 (2023) 4
91. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: CVPR (2022) 4
92. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. In: ICCV (2023) 7
93. Zheng, M., Wang, F., You, S., Qian, C., Zhang, C., Wang, X., Xu, C.: Weakly supervised contrastive learning. In: ICCV (2021) 10
94. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision (ECCV) (2022) 2
95. Zhu, X., Chen, J., Zeng, X., Liang, J., Li, C., Liu, S., Behpour, S., Xu, M.: Weakly supervised 3d semantic segmentation using cross-image consensus and inter-voxel affinity relations. In: ICCV (2021) 3