Vocabulary-Free 3D Instance Segmentation with Vision-Language Assistant

Guofeng Mei Luigi Riz Yiming Wang Fabio Poiesi
Fondazione Bruno Kessler
Via Sommarive, 18, 38123 Trento, Italy
{gmei,luriz,ywang,poiesi}@fbk.eu

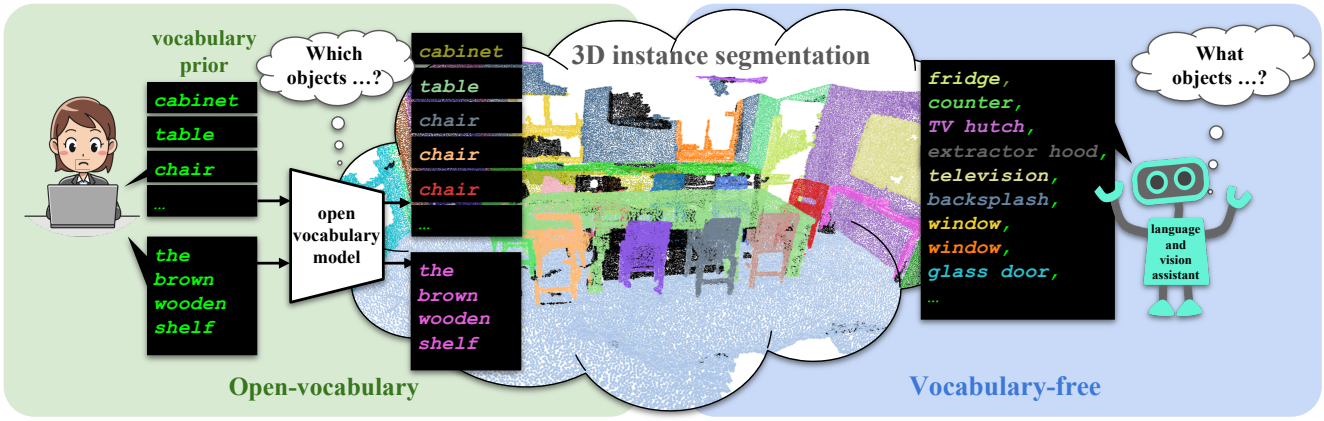

Figure 1. We introduce a vocabulary-free approach to address 3D instance segmentation that leverages vision-language assistants, moving beyond the limitations of open-vocabulary approaches. Left: 'Open-vocabulary', where 3D instances are segmented using the user-specified restricted lexical scope, *i.e.*, the 'vocabulary prior'. Right: 'Vocabulary-free', our approach can understand scenes without relying on vocabulary prior, and autonomously recognizes a wide-range of objects, *e.g.* backsplash and TV hutch.

Abstract

Most recent 3D instance segmentation methods are open vocabulary, offering a greater flexibility than closed-vocabulary methods. Yet, they are limited to reasoning within a specific set of concepts, i.e.*, the vocabulary, prompted by the user at test time. In essence, these models cannot reason in an open-ended fashion,* i.e.*, answering "*`List the objects in the scene.`*" We introduce the first method to address 3D instance segmentation in a setting that is void of any vocabulary prior, namely a vocabulary-free setting. We leverage a large vision-language assistant and an open-vocabulary 2D instance segmenter to discover and ground semantic categories on the posed images. To form 3D instance masks, we first partition the input point cloud into dense superpoints, which are then merged into 3D instance masks. We propose a novel superpoint merging strategy via spectral clustering, accounting for both mask coherence and semantic coherence that are estimated from the 2D object instance masks. We evaluate our method using ScanNet200 and Replica, outperforming existing methods in both vocabulary-free and open-vocabulary settings.* [1]

[1]Project page: https://gfmei.github.io/PoVo

1. Introduction

3D instance segmentation (3DIS) is a challenging research problem that demands instance-level semantic understanding and precise object delineation. Given a 3D scene, 3DIS aims to produce a set of binary masks associated with their semantic labels, where each of them correspond to an object instance. Traditional methods addressing 3DIS follow a closed-vocabulary paradigm [2, 29, 33, 38], where the set of semantic categories that can be encountered at test time is the same as that seen at training time. With advances in vision-language models, the 3DIS literature has rapidly evolved from closed-vocabulary methods to *open-vocabulary* methods [24, 32], where the semantic categories at test time can be different from those seen during training. Open-vocabulary 3DIS (OV3DIS) methods focus on obtaining i) instance-level 3D masks by using class-agnostic 3D segmentation methods (*e.g.* Mask3D [29] or superpoints [8]), and ii) the corresponding text-aligned mask representation by aggregating text-aligned visual representations from posed images, *e.g.* obtained by CLIP. Scene semantics are obtained by assuming certain *vocabulary prior* (left-hand side of Fig. 1): either in the form of a large finite set of categories [25, 31], akin to answer "`Choose`

among the *vocabulary*, which object do these points correspond to?", or in the form of human query with specified object instance [32], akin to answer "Do these points correspond to *an armchair with floral print*?".

What if the vocabulary prior is unknown at the inference time? Such scenarios can occur in assistive robotic applications, especially when the scene semantics are dynamically evolving, *e.g.* inclusion or replacement of instances that are not defined in the vocabulary or unknown to the user, such as an uncommon utensil or a specific painting. Thus, it would be ideal to empower 3DIS methods with the capability in answering open-ended questions, such as "What object do these points correspond to?". Inspired by the naming convention in the image recognition field [3], we term such setting as **Vo**cabulary-**F**ree **3D I**nstance **S**egmentation (**VoF3DIS**), which further extends the OV3DIS by lifting the need of any vocabulary prior. Formally, VoF3DIS aims to segment all object instances in a point cloud, with their semantic labels associated, without relying on any vocabulary prior, *e.g.* pre-defined vocabulary or user-specified query at test time (right-hand side of Fig. 1).

Previous works in such a vocabulary-free setting primarily focus on image recognition and tackle the challenge of obtaining relevant categories in a vast semantic space. Different methods have been explored to form candidate vocabularies, including retrieving from a multimodal web-scale database [3] or querying from vision-language assistants [20], *e.g.* LLaVA [18]. VoF3DIS instead focuses on understanding the 3D scene at the instance level, which presents significant new challenges, as the scale of 3D data is not comparable to that of 2D scene understanding. Moreover, it not only faces the hurdle of a vast semantic space but also demands point-level semantic representation - VoF3DIS technically nontrivial aspect, as existing large vision-language models typically process the input image as a whole [16, 18].

We propose the first point cloud vocabulary-free instance segmentation method, PoVo, which can semantically understand 3D instance masks utilizing a vision-language assistant [19]. PoVo is zero-shot and does not require any training on either 2D or 3D data. Instead, it leverages a vision-language assistant and a visual grounding model to obtain localized scene semantics from 2D posed images. Such semantic representations are then lifted to 3D to merge and refine the 3D masks that are initially segmented based on geometric features. Specifically, for each posed image, PoVo prompts the vision-language assistant to retrieve the list of objects present in the scene from the answer, forming a *scene vocabulary*. For the 3D instance proposal, PoVo first segments the 3D scene into superpoints via graph cut based on geometric features. Then we leverage a visual grounding model, *e.g.* anchored SAM [27], to obtain semantically aware object masks in posed images using the scene vocabulary. Those masks and their semantic labels are then used to guide the merging process of superpoints towards 3D instance masks via spectral clustering, using a superpoint affinity matrix based on both spatial and semantic information. Lastly, PoVo tags such 3D proposals using both vision and text information. We evaluated PoVo on two 3D scene datasets: ScanNet200 [28] and Replica [30]. To assess this newly introduced setting (VoF3DIS), we also BERT score [36] with average precision metrics to address the challenge of labeling points with concepts from a vast semantic space. Our results demonstrate that PoVo outperforms recent approaches adapted to the VoF3DIS setting. Moreover, PoVo also outperforms competitors in the open-vocabulary setting, validating its robust design in managing a large set of semantic concepts. In summary, our contributions include:

- introducing the new vocabulary-free task for 3D instance segmentation;
- proposing PoVo, the first method to transfer visually-grounded concepts identified by large vision and language models to point clouds for 3D instance segmentation;
- proposing a novel 3D mask formation strategy that merges over-segmented superpoints into instance masks using spectral clustering, considering both semantic consistency and mask coherence.

2. Related work

Vocabulary-free models. Conti et al. [3] pioneered the vocabulary-free setting, that is assigning "*an image to a class that belongs to an unconstrained language-induced semantic space at test time, without a vocabulary*". Their method, named CaSED, retrieves captions from a database that are semantically closer to the input image. From these captions, candidate categories are extracted through text parsing and filtering. CaSED estimates the similarity score between the input image and each candidate category using CLIP, leveraging both visual and textual information, to predict the best matching candidate. Subsequent works remain mostly in the image domain, *e.g.* by extending vocabulary-free image classification to semantic image segmentation [4], and exploring vision-language assistants for fine-grained image classification [20]. To the best of our knowledge, VoF3DIS is the first that extends such vocabulary-free setting to 3D instance segmentation. Instead of retrieving from a web-scale database, PoVo leverages a vision-language assistant to obtain relevant semantic concepts for the target 3D scene, making it more flexible and versatile.

Open-vocabulary 3D scene understanding. Recent advancements in 3D scene understanding mostly involve the adaptation of Vision-Language Models (VLMs) to the 3D domain, enabling semantic understanding in the open-vocabulary setting, *i.e.*, being able to recognize objects within a wide-range vocabulary, no longer being constrained

by the closed-set at training time. A common pathway is to transfer visual representation features of multi-view images to 3D points via pixel-to-point correspondences, which can then be used to either train text-aligned 3D encoders [12, 25], or to address open-vocabulary 3D scene understanding tasks directly [11]. In the following, we discuss recent open-vocabulary methods designed for 3D scene semantic segmentation and instance segmentation.

3D semantic segmentation methods aim to obtain point-level representation that are aligned with text. PLA [7] leverages image captioning models to form hierarchical 3D-caption pairs and employs contrastive learning to align point-level features with textual representation. OpenScene [25] learns a 3D network through distillation and uses open-vocabulary 2D segmentation approaches, such as OpenSeg [9] and LSeg [15], to extract pixel-level features from posed images. Differently, OV3D [12] prompts vision-language assistant to generate semantic classes with detailed descriptions from multi-view images. Text with enriched semantics are then anchored to pixels via image segments obtained with class-agnostic segmentation method, *e.g.*, SAM [13]. OV3D [12] further trains a 3D encoder to align with the pre-trained text encoder for open-vocabulary semantic segmentation. ConceptFusion [11] transfers visual features to 3D points in a training-free fashion. It employs a class-agnostic image segmentation method, like SAM [13], to localize regions containing objects in images and utilizes a vision encoder, such as CLIP [26], to generate pixel-level features, which can be lifted to 3D points for open-vocabulary scene understanding.

3D instance segmentation methods aim to obtain instance-level 3D masks with associated text-aligned per-mask representation. OpenMask3D [32] is a prior method that features open-vocabulary 3D instance segmentation. OpenMask3D relies on Mask3D [29], a class-agnostic 3D instance mask generator to generate 3D instance masks. Each 3D mask is projected to multi-view images to locate corresponding visual regions and obtain text-aligned representation with CLIP encoder. Finally, it matches per-mask representations with textual representation of user-specified queries for segmenting instances. OVIR-3D [21] processes multi-view images with an off-the-shelf open-vocabulary 2D detector, Detic [39], to produce 2D region proposals associated with text-aligned features, which are then aggregated to 3D points, forming 3D instance representation for query. Instead of using a class-agnostic 3D segmenter to generate instance masks [32], recent works [31, 35, 37] partition the 3D scene into superpoints and progressively merge them into 3D instance masks with 2D guidance. Finally, they assign a semantic label to each mask or refer to any mask using user-specified text based on CLIP. For example, SAI3D [35] builds a sparse affinity matrix that captures pairwise similarity based on the 2D masks generated by SAM to merge superpoints. Open-vocabulary 3DIS is then achieved by finding the most overlapping 3D mask with the area that are projected from 2D masks obtained by OVSeg [17] given the text query. OVSAM3D [31] project the superpoints onto 2D posed images to serve as point prompts to guide SAM for image segmentation. The image segments are then projected back to 3D for refining the 3D masks. Semantic labels are obtained from an open-vocabulary tagging method, RAM [37] with a vocabulary of about 6,400 categories. OVSAM3D also leverages a language assistant (ChatGPT) to filter out scene-irrelevant concepts. Semantic categories can then be anchored with CLIP by comparing the visual embeddings of SAM segmented 2D crops against the text embeddings of RAM-obtained tags. Open3DIS [24] further combines the class-agnostic 3D instance segmentation method, *e.g.*, Mask3D, with a 3D instance mask segmenter guided by 2D methods (*i.e.*, merging superpoints with the help of a 2D segmenter, such as SAM). For each mask, Open3DIS computes a text-aligned representation by aggregating the CLIP visual representation of multi-scale crops from multiple views.

Our work focuses on 3D instance segmentation. However, unlike previous work on open-vocabulary 3D instance segmentation, VoF3DIS features a novel setting that operates under the assumption that target classes are unknown during inference. PoVo addresses VoF3DIS in a training-free manner by leveraging vision-language assistants to provide scene semantics and ground them into 3D instance segments.

3. Vocabulary-free 3D scene understanding

Definition. Vocabulary-free 3D instance segmentation (VoF3DIS) aims to assign a semantic label to each 3D instance mask in a point cloud without relying on any predefined list of categories (vocabulary) at test time. Formally, given a point cloud $\mathcal{P} = \{\mathbf{p}\}$, where $\mathbf{p} \in \mathbb{R}^d$ s.t. $d \geq 3$[2], The point cloud is decomposed into a set of 3D instance masks $\mathcal{M}^{3D}$, where each mask $M_i^{3D} \in \mathcal{M}^{3D}$ is a set of binary values with ones indicating its corresponding points belonging to the i^{th} object instance, and zeros otherwise. VoF3DIS involves assigning a class $c \in \mathcal{S}$ to each 3D instance mask M_i^{3D}, where $\mathcal{S}$ represents an unconstrained semantic space. For example, BabelNet [23] contains millions of semantic concepts, that is four magnitudes larger than the semantic classes annotated in ScanNet200 [28]. The objective is to design a function f that maps 3D masks to concepts, formally defined as $f : \mathcal{M}^{3D} \rightarrow \mathcal{S}$. At test time, the function f has access to the point cloud $\mathcal{P}$ and to a source that provides vast semantic concepts approximating $\mathcal{S}$. Potential semantic sources, as discussed in [3], can be either in the format of a web-scale database or a model that is trained with such database. VoF3DIS requires searching for relevant concepts

[2] $d = 3$ represents the basic case in which a point is represented by a 3D coordinate (LiDAR capture). In some instances, $d = 4$ or $d = 6$ if LiDAR luminance, or RGB information are available, respectively.

from a vast semantic source and requires point-level semantic grounding, making it significantly more challenging than image classification as in [3].

Challenges. A key challenge in distinguishing concepts within a vast semantic space in point clouds is ensuring spatial consistency when assigning labels to points. This requires models to not only label individual points accurately but also to guarantee that adjacent points, likely belonging to the same object, are assigned to the same label. Moreover, point sparsity may result in incomplete or ambiguous object representations, making it difficult to distinguish small or fine-grained objects. Additionally, reconstruction noise can render certain elements of the scene geometrically indistinguishable, making it challenging to differentiate between foreground and background or objects with similar shapes but different functions, particularly in cases where photometric information is missing or inaccurate.

4. Our approach

Given the point cloud $\mathcal{P}$ of a 3D scene and the corresponding set of N posed images $\mathcal{V} = \{I_n\}_{n=1}^{N}$, the proposed PoVo predicts 3D instance masks with their associated semantic labels without knowing a predefined vocabulary. As shown in Fig. 2, PoVo first utilizes a large vision-language assistant and an open-vocabulary 2D instance segmentation model to identify and ground objects on each posed image I_n, forming the scene vocabulary $\mathcal{C}$.

Meanwhile, we partition the 3D scene $\mathcal{P}$ into geometrically-coherent superpoints $\mathcal{Q}$, to serve as initial seeds for 3D instance proposals. Then, with the semantic-aware instance masks from multi-view images, we propose a novel procedure in representing superpoints and guiding their merging into 3D instance masks, using both the grounded semantic labels and their instance masks. Specifically, by projecting each 3D superpoint onto image planes, and checking its overlapping with 2D instance masks, we can aggregate the semantic labels from multiple views within each superpoint. Once each superpoint is associated to a semantic label, we then perform superpoint merging to form 3D instance masks via spectral clustering. To do so, we define an affinity matrix among superpoints constructed by both mask coherence scores computed with the 2D instance masks, and semantic coherence scores computed with the per-superpoint textual embeddings. Finally, for each 3D instance proposal, we obtain the text-aligned representation by aggregating the CLIP visual representation of multi-scale object crops from multi-view images (as in [24]). We further enrich such vision-based representation with textual representation coming from the merged superpoints. The text-aligned mask representation enables the semantic assignment to instance masks with the scene vocabulary $\mathcal{C}$.

It is worth to mention that PoVo can not only address VoF3DIS, but it is also compatible with the open-vocabulary setting, by performing semantic assignment with any given vocabulary or user-specified prompt.

4.1. Scene vocabulary generation

PoVo first utilizes a large vision-language assistant to identify the scene vocabulary $\mathcal{C}$, *i.e.*, the list of object categories in the scene, that are grounded in the multi-view images. Specifically, for each posed image I_n, we prompt the vision-language assistant with "`List the object names in the scene`". We then parse the response to obtain the list of objects, $\mathcal{C}_n^-$, present in the image from the answer. To mitigate the potential hallucination of object presence by the vision-language assistant, we subsequently employ an open-vocabulary 2D instance segmentation model, *e.g.* grounded SAM, to ground all categories in $\mathcal{C}_n^-$, obtaining the grounded object categories $\mathcal{C}_n$, as well as the set of masks $\mathcal{M}_n^{2D}$ for each object on the 2D posed image I_n.

With $\mathcal{C}_n$, we then construct the scene vocabulary $\mathcal{C}$ for each point cloud by retaining only the unique categories from the combined sets of $\mathcal{C}_n$, formally as $\mathcal{C} = \bigcup_{n=1}^{N} \mathcal{C}_n$, where $\bigcup$ denotes the union operation.

The grounded object categories $\mathcal{C}_n$, and their corresponding instance masks $\mathcal{M}_n^{2D}$ on each 2D posed image, are further exploited in the process of representing and merging superpoints towards 3D instance masks $\mathcal{M}^{3D}$, as detailed in the following section.

4.2. 3D instance mask formation

We leverage geometrically coherent over-segmented superpoints to initialize 3D instance formation. In addition to being a common practice in prior works [24, 31, 35], superpoint initialization is more generalizable and better suited for zero-shot setups compared to pre-trained class-agnostic 3D instance segmentation models, such as Mask3D [29], which are trained on datasets used for method evaluation. In the following, we detail the process of generating superpoints and how they are merged into 3D instances by leveraging the results of instance segmentation on posed images.

Superpoint generation. We use graph cut to group points into geometrically homogeneous regions, yielding a set of M superpoints $\mathcal{Q} = \{Q_i\}_{i=1}^{M}$, where Q_i is a binary mask of points in $\mathcal{P}$. Superpoints are dense partitions of the 3D scene. Neighboring superpoints are likely corresponding to the same semantic label. For example, a table, according to geometric features, might be partitioned into multiple superpoints corresponding to different surface planes, with each plane sharing the same semantic label, *i.e.*, "Table". Via merging superpoints, we can form semantically-coherent 3D instance masks.

Superpoint merging by spectral clustering. We aim to merge superpoints that tend to overlap with the same 2D masks when projected onto their respective images while ensuring semantic consistency. To this end, we define: i)

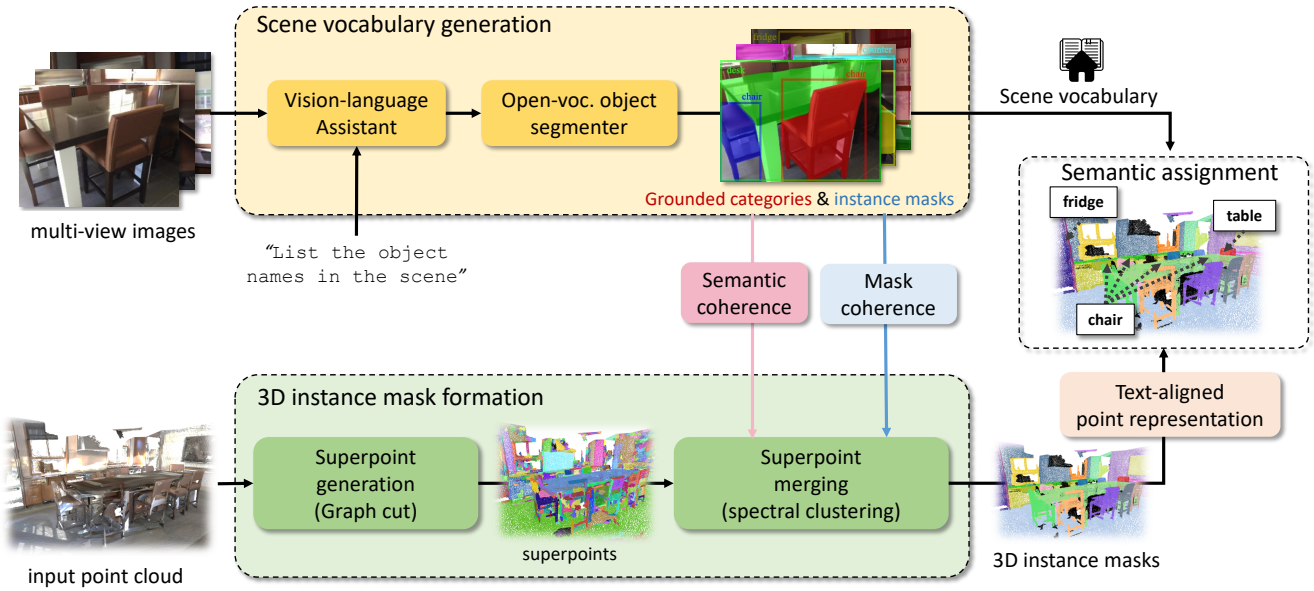

Figure 2. PoVo's architecture to address the VoF3DIS task. To generate the scene vocabulary, we start from multi-view images and use a vision-language assistant to identify lists of objects contained in each posed image. Then, we run an open-vocabulary object segmenter to ground the identified categories with instance masks to mitigate the potential risk of hallucination. We finally obtain the scene vocabulary, *i.e.*, the list of unique grounded categories among all posed images. In parallel, to form 3D instance masks, we extract superpoints from the point cloud with graph cut. We then merge those dense superpoints to form 3D instance masks, considering both the *semantic coherence* and *mask coherence* which are computed with the 2D object masks. Finally, we obtain text-aligned point features for all points within each 3D instance mask, which are then used to assign the semantic category within the scene vocabulary. In addition to VoF3DIS, PoVo is also able to deal with the open-vocabulary setting, by substituting the scene vocabulary with any predefined vocabulary or user-specified prompt.

a *mask coherence* score $a^M_{i,j}$ that quantifies the likelihood that two superpoints, Q_i and Q_j, belong to the same object instance; ii) a *semantic coherence* score $a^S_{i,j}$ that quantifies the probability that two superpoints correspond to the same class; iii) a *spatial connectivity* score $a^C_{i,j}$ that indicates adjacency if the distance between the superpoints falls within a predefined threshold.

We derive the mask coherence score by evaluating the overlap ratio between the 3D superpoints Q_i and Q_j when projected onto the image plane of a 2D mask. Specifically, for each 2D instance mask $M^{2D}_t \in \{\mathcal{M}^{2D}_n\}^N_{n=1}$, we calculate the Intersection over Union (IoU) $O_{i,t}$ between each superpoint Q_i and the image pixels corresponding to Q_i. This is done by projecting all points of Q_i onto the image plane of M^{2D}_t using the known camera matrix, while excluding points outside the camera's field of view. A superpoint is considered to have sufficient overlap with a 2D mask if the IoU is higher than a threshold, *i.e.*, $O_{i,t} > \tau_{iou}$. We then compute the mask coherence score a^M_{ij} between two superpoints Q_i and Q_j, accounting all 2D instance masks as:

$$a^M_{ij} = \sum_t g\left(O_{i,t}, \tau_{iou}\right) \cdot g\left(O_{j,t}, \tau_{iou}\right), \qquad (1)$$

where T is the total number of 2D instance masks and the function $g(x, \tau)$ is defined as $g(x, \tau) = x$ if $x > \tau$, and 0 otherwise. By computing a^M_{ij} among all pairs of superpoints, we obtain the mask coherence matrix A^M.

To compute the semantic coherence score a^S_{ij}, we first obtain the semantic representation per superpoint. Specifically, for each superpoint Q_i, we identify the top K 2D masks based on the IoU $O_{i,t}$ and their corresponding semantic labels. The most frequent label among these 2D masks is assigned as the semantic label for the superpoint Q_i. We then obtain its semantic representation by encoding the label text of Q_i into a feature vector f_{Q_i} using a text encoder. We then calculate the cosine similarity using the semantic representation between Q_i and Q_j as a^S_{ij}:

$$a^S_{ij} = \frac{f^\top_{Q_i} f_{Q_j}}{\|f_{Q_i}\| \|f_{Q_j}\|} \odot \left(f^\top_{Q_i} f_{Q_j} > \tau_{sim}\right). \qquad (2)$$

Similarly, we obtain the semantic coherence matrix A^S by computing a^S_{ij} among all pairs of superpoints.

To efficiently compute the spatial connectivity scores denoted as an adjacency matrix A^C, we sample points K=64 (using the farthest point sampling) from each superpoint and set the threshold as twice the average distance τ_c of the sampled points. The distance between two superpoints Q_i and Q_j is defined as $d(Q_i, Q_j) = \min_{p \in Q_i, q \in Q_j} \|p - q\|$. The adjacency score a^C_{ij} is then assigned as a^C_{ij}=1 if $d(Q_i, Q_j) < \tau_c$ and $a^C_{ij} = 0$ otherwise.

With A^M, A^S, and A^C, we construct an affinity matrix A for the superpoints $\mathcal{Q}$ satisfying $A = A^M \odot A^S \odot A^C$. Next, we compute the Laplacian eigenvec-

tors of A, where the Laplacian matrix is defined as $L{=}D^{-1/2}(D{-}A)D^{-1/2}$. We then merge superpoints using its eigenvectors: $\{y_0,\ldots,y_{M-1}\}{=}\text{eigs}(L)$, with the corresponding eigenvalues $\{\lambda_0,\ldots,\lambda_{M-1}\}$ ranked in ascending order, where $\lambda_0 = 0$. We discretize the first H eigenvectors $\{y_0,\ldots,y_H\}$ of L by clustering them across the eigenvector dimension using K-means. H is determined by eigengap heuristic as $H{=}\underset{1\leq j\leq M-2}{\text{argmax}}\ (\lambda_{j+1}-\lambda_j)$. In doing so, we can extend Q_i with neighboring superpoints Q_j that meet the overlapping condition (τ_{iou}) with the highest semantic similarity. Please refer to Sec. A of the supplementary materials for more details.

4.3. Text-aligned point representation

Unlike prior works [11, 25], which leverage only the visual encoder of pre-trained vision-language models such as CLIP, we employ both the vision encoder and the text encoder for per-point representation to mitigate the potential modality gap observed in [3]. Specifically, given a point $\mathbf{p}_l$ and its corresponding superpoint Q_i, we first use the vision feature extraction method provided by Open3DIS [24] to extract the per-point feature f_l^v. The CLIP vision encoder is then used to encode image crops from multiple posed images at multiple scales, obtained by projecting their corresponding 3D masks onto the posed images. Finally, we enrich f_l^v with the superpoint-level feature f_{Q_i}, obtaining the final point-level feature f_l via mean pooling.

5. Experiments

We evaluate PoVo on the 3D scene instance segmentation task in open-vocabulary and vocabulary-free settings. We compare PoVo with state-of-the-art methods by using two common benchmark datasets. We present quantitative and qualitative results and ablation studies.

Datasets. We quantitatively evaluate PoVo by using 3D scans of real scenes from the ScanNet200 [28] and Replica [30] datasets. These datasets include both instance and semantic (vocabulary) annotations. ScanNet200 [28] contains a validation set of 312 indoor scans with 200 object categories, which is significantly more than its predecessor, that is ScanNet [5], which features 20 semantic classes only. Replica [30] consists of 8 evaluation scenes with 48 classes. In the supplementary material, we also present additional results obtained on the S3DIS dataset [1] (Sec. B).

Performance metrics. Following the experimental setup of ScanNet [5], we compute the score at average mask thresholds ranging from 50% to 95% in 5% increments as the Average Precision (AP). We then compute the AP at specific mask overlap thresholds of 50% and 25% as AP50 and AP_{25}, respectively. For ScanNet200, we report results for different category groups such as AP_{head}, AP_{com}, and AP_{tail} [28].

In the open-vocabulary setting, evaluation is straightforward as the vocabulary provided by the underlying dataset can be used directly, however in VoF3DIS the vision-language assistant can label objects differently than those labeled by humans in the ground truth [3]. We mitigate this problem by using the BERT Similarity [36] that can quantify the semantic relevance of the predicted point label in relation to the ground-truth label, as in [14]. The BERT Similarity is 1 when the similarity between predicted and ground-truth labels is highest. We use a stringent threshold $\tau_{bert} = 0.8$ on this similarity to deem a predicted label correct.

Baselines. We compare PoVo with state-of-the-art methods, including OpenScene [25], OpenMask3D [32], OVIR-3D [21], SAM3D [34], SAI3D [35], OVSAM3D [31], and Open3DIS [24]. OpenScene is adapted for instance segmentation using Mask3D [29] as in [24]: we name this version OpenScene*. Since there are no existing scene understanding methods specifically designed for the VoF3DIS setting (SAM mask + vocabulary-free semantics), we implemented a set of baselines using state-of-the-art methods. We adapt the open-vocabulary instance scene segmentation methods Open3DIS [24] and SAM3D [34] to the VoF3DIS setting, and name these versions Open3DIS† and SAM3D†, respectively. We only use the 2D mask proposals provided by Open3DIS to be comparable with our method. We replace the user-provided vocabulary, *i.e.*, the full category list of each dataset, with the one generated by LLaVA as described in Sect. 4.1. Lastly, we evaluate PoVo in the open-vocabulary setting for further comparison of PoVo with state-of-the-art methods, *i.e.*, 2D/3D mask + Open-vocab. semantic.

Implementation Details. PoVo is implemented with PyTorch using the original implementations of CLIP [26], LLaVA [18], and Grounded-SAM[3]. For LLaVA, we use llava-v1.6-mistral-7b, while for CLIP, we use ViT-L/14. We set $\tau_{iou} = 0.9, \tau_{sim} = 0.9$ for all experiments. For each superpoint, we choose top $K = 5$ view masks with the largest IoU of projected points. Experiments are run on a single NVIDIA A40 48GB RAM. We use the original source codes for the baselines.

5.1. Analysis of the results

ScanNet200. Tab. 1 reports OV3DIS and VoF3DIS results on the ScanNet200 dataset. Following [10, 23], we test both the OV3DIS and VoF3DIS setting in the validation set. The first and second sections of the table compare PoVo adapted to the open-vocabulary setting (3D/2D mask Open-vocab. semantic) and baselines. Although not explicitly designed for an open vocabulary setting, PoVo outperforms the other baselines. A significant distinction between PoVo and Open3DIS lies in the processing of multi-view images. PoVo retrieves concepts through an assistant and transfers their features to the 3D points, whereas Open3DIS pre-processes images using open-vocabulary segmentation and transfers visual foundational features to the 3D points. We observed that

[3] https://github.com/IDEA-Research/Grounded-Segment-Anything

Table 1. 3D instance segmentation results on ScanNet200. Best result for each metric is in **bold**.

Method	Semantic	AP	AP_{50}	AP_{25}	AP_{head}	AP_{com}	AP_{tail}
3D mask + Open-vocab. semantic							
OpenScene* [25]	OpenSeg [9]	11.7	15.2	17.8	13.4	11.6	9.9
OpenMask3D [32]	CLIP [26]	15.4	19.9	23.1	17.1	14.1	14.9
2D mask + Open-vocab. semantic							
OVIR-3D [21]	Detic [39]	9.3	18.7	25.0	9.8	9.4	8.5
SAM3D [34]	OpenSeg [9]	7.4	11.2	16.2	6.7	8.0	7.6
SAI3D [35]	OpenSeg [9]	9.6	14.7	19.0	9.2	10.5	9.1
OVSAM3D [31]	CLIP [26]	9.0	13.6	19.4	9.1	7.5	10.8
Open3DIS [24]	CLIP [26]	18.2	26.1	31.4	18.9	16.5	19.2
PoVo	CLIP [26]	**22.4**	**27.9**	**34.4**	**20.7**	**20.2**	**20.6**
2D mask + Vocab.-free semantic							
SAM3D† [34]	OpenSeg [9]	6.7	10.4	15.7	6.2	7.4	6.8
Open3DIS† [24]	CLIP [26]	17.6	25.4	30.9	18.4	15.8	18.2
PoVo	CLIP [26]	**21.6**	**26.7**	**33.0**	**19.5**	**19.1**	**19.4**

the segmentation performance of Open3DIS is limited by this preprocessing and that the resulting image segmentation is relatively noisy. In contrast, our method transfers and aggregates concepts based on vision and text features, followed by superpoint-based pooling to mitigate noisy features, thereby enhancing robustness. The third section of the table reports the results in the VoF3DIS setting, where we compare PoVo with Open3DIS† and SAM3D†. When Open3DIS† is provided with the list of objects identified by LLaVA, its performance is inferior to that in the open-vocabulary setting. This suggests that Open3DIS† struggles to handle a large corpus of concepts since it only considers vision features. By contrast, PoVo significantly outperforms these baselines. Specifically, PoVo achieves strong and stable performance across both common and rare class categories, as shown by the metrics AP_{head}, AP_{com}, and AP_{tail}. This is because our method is not trained on 3D annotated data, but instead leverages transferred semantic information from large vision-language models and 2D foundational models, which have been trained on massive 2D datasets and exposed to a wide range of categories. Fig. 3 shows the qualitative results of text-driven 3D instance segmentation. In the first row, we observe that PoVo can accurately segment most parts of the scene with the correct labels. Our model can successfully segment instances based on various types of input text prompts, which include object categories not present in the predefined labels, objects' functionality, branch, and other properties. In the second row, we have highlighted the objects in the corresponding RGB images with boxes. Sec. B of the Supplementary Material analyses more results.

Replica. Tab. 2 reports open-vocabulary and vocabulary-free results on the Replica dataset. PoVo outperforms all the other baselines on both the open-vocabulary and VoF3DIS settings. Specifically, in the former, our approach outperforms Open3DIS [24] and OVIR-3D [21] by margins of +2.7 and +9.7 in AP, respectively. In the latter, PoVo outperforms Open3DIS† [24] by a margin of +1.6 in AP. This performance gap highlights the effectiveness of our approach in handling unseen categories, bolstered by the 2D foundation model and vision-language model assistance.

Table 2. 3D instance segmentation results on Replica. Best result for each metric is in **bold**.

Method	AP	AP_{50}	AP_{25}
3D mask + Open-vocab. semantic			
OpenScene* [25]	10.9	15.6	17.3
OpenMask3D [32]	13.1	18.4	24.2
2D mask + Open-vocab. semantic			
OVIR-3D [21]	11.1	20.5	27.5
Open3DIS [24]	18.1	26.7	30.5
PoVo	**20.8**	**28.7**	**34.4**
2D mask + Vocab.-free semantic			
Open3DIS† [24]	17.3	25.8	29.0
PoVo	**18.9**	**27.6**	**31.9**

5.2. Ablation study

To evaluate the effectiveness of our model design, we conducted a series of ablation studies on the validation set of ScanNet200. More ablation studies are given in the Supplementary Material.

How effective are superpoints to guide mask representation? There are two key technical designs that we consider important for PoVo: text-embedding-enhanced features and superpoint-based average pooling. Tab. 3 shows the performance of various feature fusion strategies. The first row displays the results using only vision features, denoted as VisEmb. We use the CLIP visual encoder to obtain visual representations for each 3D proposal by aggregating information from multiple views. While VisEmb outperforms using only text representations (TxtEmb), their fusion enables

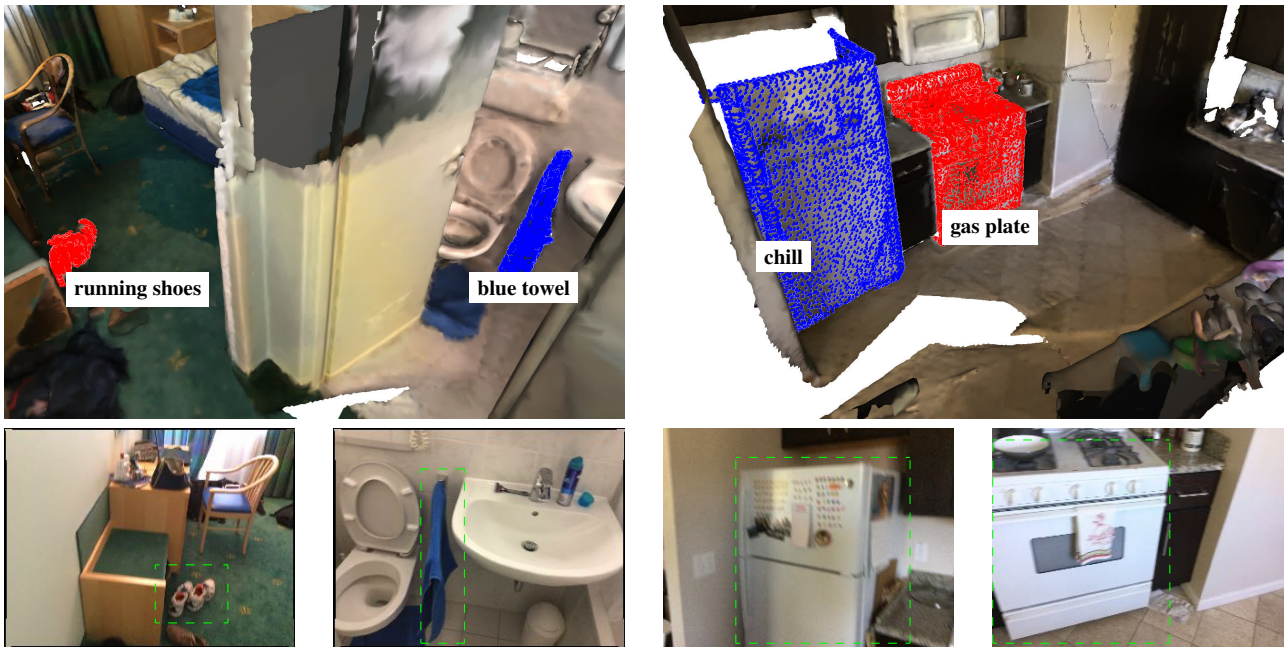

Figure 3. Qualitative results obtained by PoVo in the VoF3DIS setting on ScanNet200. Instance masks are generated by querying PoVo with query vocabulary. The instance with the highest similarity score to the query's embedding is highlighted in the point clouds. Green boxes outline the regions of the objects in the corresponding RGB images.

Table 3. Ablation study of PoVo for instance feature extraction on ScanNet200 dataset. Each variant of PoVo has one component that differs from the final version of PoVo.

Setting	**AP**	**AP$_{50}$**	**AP$_{25}$**	**AP$_{head}$**	**AP$_{com}$**	**AP$_{tail}$**
VisEmb	20.4	25.8	32.3	18.6	18.1	18.2
TxtEmb	15.3	22.3	28.6	17.4	17.3	17.6
w/o SPool	20.3	25.6	31.9	18.4	18.0	18.2
PoVo	**21.6**	**26.7**	**33.0**	**19.5**	**19.1**	**19.4**

Table 4. Ablation study of PoVo using text embedding for superpoint merging and superpoint-based pooling on ScanNet200.

Setting	**AP**	**AP$_{50}$**	**AP$_{25}$**	**AP$_{head}$**	**AP$_{com}$**	**AP$_{tail}$**
w/o TxtSim	19.8	25.1	30.9	17.7	17.3	17.2
PoVo	**21.6**	**26.7**	**33.0**	**19.5**	**19.1**	**19.4**

PoVo to achieve overall and consistently improved results. Since the generated point-level features may be noisy, we introduce a feature refinement module that averages representations (pooling) at the superpoint level, assuming that points belonging to the same superpoint should share the same features. The third row (w/o SPool) shows the results without superpoint-based pooling, which are inferior to those of PoVo across all metrics.

How effective is text embedding superpoint merging? We experimentally assess the influence of text feature similarity (TxtSim) guided merging on the performance of PoVo. Second row of Tab. 4 shows that incorporating text embedding for superpoint merging can enhance instance segmentation performance. This improvement is achieved because 2D masks can encompass background regions or nearby objects, rendering IoU alone insufficient for accurately determining the association of superpoints with a 3D proposal. Leveraging text feature similarity helps PoVo to mitigate this issue.

6. Conclusions

We presented a novel approach to 3D scene understanding that operates without the need for a predefined vocabulary. By integrating a large vision-language assistant with an open-vocabulary 2D instance segmenter, our PoVo can autonomously identify and label each 3D instance in a scene. Furthermore, our innovative use of superpoints, in conjunction with spectral clustering, enables our system to generate robust 3D instance proposals. We evaluated PoVo on two point cloud datasets, ScanNet200 and Replica, and showed that PoVo outperforms recent approaches adapted to the VoF3DIS setting, as well as in the open vocabulary setting.

Because PoVo can effectively exploit vision-language assistant understanding with point cloud data in a training-free manner, an exciting future research direction includes exploring new 3D scene understanding tasks such as affordances, using the soon-to-be-released dataset SceneFun3D [6]. Moreover, our goal is to improve geometric understanding using the most recent zero-shot approaches, such as [22]. Lastly, implementing new large vision-language assistants can be a viable way to enhance point cloud understanding even further.

Acknowledgment. This work was sponsored by PNRR ICSC National Research Centre for HPC, Big Data and Quantum Computing (CN00000013) and the FAIR - Future AI Research (PE00000013), funded by NextGeneration EU.

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. 6

[2] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *ICCV*, pages 15467–15476, 2021. 1

[3] A. Conti et al. Vocabulary-free Image Classification. In *NeurIPS*, 2023. 2, 3, 4, 6

[4] A. Conti et al. Vocabulary-free image classification and semantic segmentation. *arXiv:2404.10864v1*, 2024. 2

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 6

[6] Alexandros Delitzas, Ayça Takmaz, Robert Sumner, Federico Tombari, Marc Pollefeys, and Francis Engelmann. SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes. In *CVPR*, 2024. 8

[7] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. PLA: Language-Driven Open-Vocabulary 3D Scene Understanding. In *CVPR*, 2023. 3

[8] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59:167–181, 2004. 1

[9] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022. 3, 7

[10] Ji Hou, Xiaoliang Dai, Zijian He, Angela Dai, and Matthias Nießner. Mask3d: Pre-training 2d vision transformers by learning masked 3d priors. In *CVPR*, pages 13510–13519, 2023. 6

[11] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, et al. ConceptFusion: Open-set multimodal 3D mapping. In *RSS*, 2023. 3, 6

[12] Li Jiang, Shaoshuai Shi, and Bernt Schiele. Open-vocabulary 3d semantic segmentation with foundation models. In *CVPR*, pages 21284–21294, 2024. 3

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything, 2023. arXiv:2304.02643. 3

[14] S. Koch et al. Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In *CVPR*, 2024. 6

[15] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 3

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2

[17] 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 *CVPR*, pages 7061–7070, 2023. 3

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 6

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 2

[20] Mingxuan Liu, Subhankar Roy, Wenjing Li, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Democratizing fine-grained visual recognition with large language models. In *ICLR*, 2024. 2

[21] Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, and Kostas Bekris. OVIR-3D: Open-vocabulary 3d instance retrieval without training on 3d data. In *CoRL*, pages 1610–1620. PMLR, 2023. 3, 6, 7

[22] Guofeng Mei, Luigi Riz, Yiming Wang, and Fabio Poiesi. Geometrically-driven aggregation for zero-shot 3d point cloud understanding. In *CVPR*, 2024. 8

[23] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: Building a very large multilingual semantic network. In *ACL*, 2010. 3, 6

[24] Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *CVPR*, pages 4018–4028, 2024. 1, 3, 4, 6, 7

[25] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, 2023. 1, 3, 6, 7

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 6, 7

[27] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 2

[28] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *ECCV*, 2022. 2, 3, 6

[29] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *ICRA*, pages 8216–8223. IEEE, 2023. 1, 3, 4, 6

[30] 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. 2, 6

[31] Hanchen Tai, Qingdong He, Jiangning Zhang, Yijie Qian, Zhenyu Zhang, Xiaobin Hu, Yabiao Wang, and Yong Liu. Open-vocabulary sam3d: Understand any 3d scene. *arXiv preprint arXiv:2405.15580*, 2024. 1, 3, 4, 6, 7

[32] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *NeurIPS*, 2023. 1, 2, 3, 6, 7

[33] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *CVPR*, 2022. 1

[34] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023. 6, 7

[35] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *CVPR*, pages 3292–3302, 2024. 3, 4, 6, 7

[36] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *ICLR*, 2020. 2, 6

[37] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *CVPR*, pages 1724–1732, 2024. 3

[38] Min Zhong, Xinghao Chen, Xiaokang Chen, Gang Zeng, and Yunhe Wang. Maskgroup: Hierarchical point grouping and masking for 3d instance segmentation. In *ICME*, pages 1–6. IEEE, 2022. 1

[39] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 3, 7