SAMPro3D: Locating SAM Prompts in 3D for Zero-Shot Instance Segmentation

Mutian Xu[1] Xingyilang Yin[1] Lingteng Qiu[1] Yang Liu[3] Xin Tong[3] Xiaoguang Han[1,2*]

[1]SSE, CUHKSZ [2]FNii, CUHKSZ [3]Microsoft Research Asia

mutianxu.github.io/sampro3d

Abstract

We introduce SAMPro3D for zero-shot instance segmentation of 3D scenes. Given the 3D point cloud and multiple posed RGB-D frames of 3D scenes, our approach segments 3D instances by applying the pretrained Segment Anything Model (SAM) to 2D frames. Our key idea involves locating SAM prompts in 3D to align their projected pixel prompts across frames, ensuring the view consistency of SAM-predicted masks. Moreover, we suggest selecting prompts from the initial set guided by the information of SAM-predicted masks across all views, which enhances the overall performance. We further propose to consolidate different prompts if they are segmenting different surface parts of the same 3D instance, bringing a more comprehensive segmentation. Notably, our method does ***not*** *require any additional training. Extensive experiments on diverse benchmarks show that our method achieves comparable or better performance compared to previous zero-shot or fully supervised approaches, and in many cases surpasses human annotations. Furthermore, since our fine-grained predictions often lack annotations in available datasets, we present ScanNet200-Fine50 test data which provides fine-grained annotations on 50 scenes from ScanNet200 dataset.*

1. Introduction

Instance segmentation of 3D scenes plays a vital role in diverse applications such as augmented reality, room navigation, and autonomous driving. The objective is to predict 3D instance masks from input 3D scenes which are often represented by meshes, point clouds, and posed RGB-D images. Traditional methods for this task [12, 23, 31, 62, 67, 74, 75, 85] lack the *zero-shot* capability. They often struggle to accurately segment newly introduced object categories that were not encountered during training [34, 47]. Despite recent efforts [4, 7, 15–17, 22, 27, 29, 40, 50, 52, 66] that harness vision foundation models [3, 32, 56] to enhance zero-shot 3D scene segmentation, they necessitate either 3D pretrained networks or training on domain-specific data. As a result, directly applying them to novel 3D scenes remains challenging in terms of generalization.

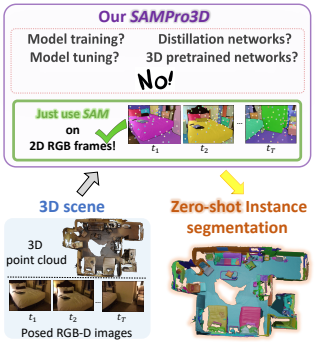

Figure 1. We introduce *SAMPro3D* for *zero-shot* instance segmentation of 3D scenes. Given the 3D point cloud and posed RGB-D frames of 3D scenes, our approach uses the Segment Anything Model (SAM) [32] on RGB frames to segment 3D instances. Our method does **not** require additional training on domain-specific data. See Fig. 5 for more impressive results.

In the field of 2D image segmentation, the Segment Anything Model (SAM) [32] brought the breakthrough. Trained on an extensive SA-1B dataset [32], SAM can "segment any unfamiliar images" without further training, by accepting various input prompts that specify where or what is to be segmented in an image. The final stage of SAM (referred to as automatic-SAM) automatically generates prompts and corresponding segmentations on the image, where each single prompt accurately segments one 2D instance. Having witnessed the great power of SAM, and recognizing that a 3D scene is essentially a combination of multiple 2D views, a curious **question** arises: Given 3D scene point clouds with posed RGB-D frames, is it possible to apply SAM *directly* to 2D frames for zero-shot 3D instance segmentation *without* additional training?

There are several potential attempts to explore this question: (1) A recent project called SAM3D [81] utilizes automatic-SAM to individual 2D frames. The resulting 2D masks are then projected into 3D space and iteratively fused

*Corresponding author

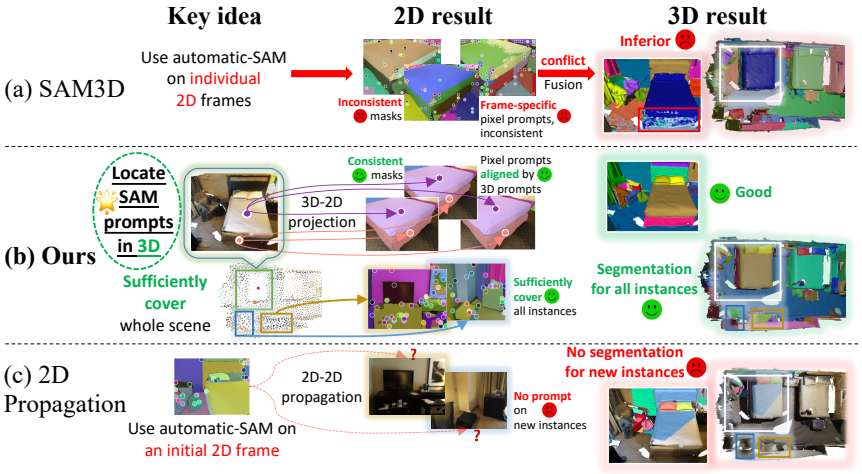

Figure 2. **The comparison of our key idea and others**. Our method (b) *locates SAM prompts in **3D***, which aligns pixel prompts across frames, bringing the frame consistency of prompts and their masks, and can handle newly emerged instances. Here we use random colors to visualize 3D results for instance discrimination, so there is no correlation between the colors assigned to 2D and 3D instances.

to obtain the final result. However, SAM3D assigns *frame-specific* pixel prompts that lack view consistency, producing inconsistent 2D masks across frames. As a result, fusing these masks may cause substantial conflicts in the segmentation of the same area, yielding inferior 3D results (see Fig. 2 (a)). (2) To realize prompt consistency across frames, one possible solution is to employ automatic-SAM on an initial frame to generate 2D-pixel prompts, which can then be propagated to subsequent frames, analogous to SAM-PT [57] for video tracking. Nevertheless, while videos processed by SAM-PT typically involve foreground objects that consistently appear across all frames, 3D scenes pose a different challenge. In 3D scenes, instances that initially appear may disappear in subsequent frames, and new instances may emerge. Consequently, initial prompts cannot be propagated to cover *newly emerged* instances in other frames of 3D scenes, resulting in the absence of segmentation masks for these instances. (see Fig. 2 (c)).

To tackle the aforementioned challenges, we present a simple yet effective method called SAMPro3D. The key idea of SAMPro3D is to *locate SAM prompts in **3D*** scene point clouds and project these 3D prompts onto 2D frames to get pixel prompts for using SAM (see Fig. 2 (b)). In this way, a 3D input point serves as the natural prompt to align the pixel prompts projected from this 3D point across different frames, making both pixel prompts and their SAM-predicted masks for the same 3D instance exhibit consistency across frames. Moreover, as our 3D prompts sufficiently cover all instances, we can obtain 3D segmentations for all instances in the scene.

Building upon this key idea, our framework is designed in a bottom-up manner. It begins by sampling 3D points from the input scene as the initial SAM prompts. Subsequently, we introduce a novel View-Guided Prompt Selection algorithm to select these initial prompts. It examines the quality of the segmentation masks generated by initial prompts within each view and accumulates the examinations across *all views*. By doing so, it selects and retains high-quality prompts for segmenting each instance, thereby enhancing the overall performance. However, we find that in some cases, a single retained prompt may only segment part of a 3D instance due to the limited visible field of 2D camera views. To address this, we further design a Surface-Based Prompt Consolidation strategy to consolidate 3D prompts that exhibit certain intersections in their 3D masked *surfaces* into one single prompt, as they are likely segmenting different parts of the same 3D instance. This brings a more comprehensive segmentation of 3D instances. Finally, we project all input points of the scene onto each segmented frame and accumulate their predictions across frames to derive the final 3D segmentation.

Our underlying design logic is to build an automatic-SAM tailored for 3D instance segmentation. We aim to *automatically* generate SAM prompts and ensure the consistency of their 2D segmentation for the same 3D instance across different frames, ultimately achieving that every *single* prompt accurately segments *one* 3D instance. Notably, our method does **not** require additional training or 3D pre-trained networks on domain-specific data. This preserves SAM's zero-shot ability and enables future applications to directly segment new 3D scenes without the need to gather plenty of training data.

We conduct extensive experiments on diverse benchmarks, including ScanNet [13] containing indoor scenes, ScanNet200 [59] providing more comprehensive annotations, ScanNet++ [82] offering more detailed segmentation masks, and KITTI-360 [37] featuring outdoor suburban scenes, where our approach demonstrates high effectiveness. In addition, we observed that our method often segments fine-grained instances that lack annotations

in available datasets. For better evaluation, we present *ScanNet200-Fine50* test data, providing more fine-grained annotations of 50 scenes from ScanNet200 validation set.

Our contributions are summarized as:

- To the best of our knowledge, we are the **first** to set SAM prompts on *3D* surfaces and *scatter* prompts to 2D views for 3D segmentation. This ensures the identity consistency of the 2D prompts across frames and can deal with emerging instances in other frames.
- Based on this key idea, we propose SAMPro3D for zero-shot 3D instance segmentation, equipped with novel prompt selection and consolidation that effectively enhance segmentation quality and comprehensiveness.
- Rich experiments show that our method consistently achieves higher quality and more diverse segmentations than previous zero-shot or fully supervised approaches, and in many cases surpasses human-level annotations.
- We present ScanNet200-Fine50 test data with more fine-grained annotations.

2. Related Work

Closed-set 3D scene understanding. The field of 3D scene understanding has been dominated by closed set methods which primarily focus on training deep neural networks on domain-specific datasets [1, 2, 5, 13, 19, 25, 36, 64, 65, 76]. The first line of research focuses on improving representation learning from human-annotated 3D labels [12, 24, 35, 38, 41, 44, 53, 54, 71, 73, 75, 79, 85], for solving different 3D scene understanding tasks [6, 9, 21, 30, 31, 42, 49, 55, 62, 63]. Another stream of works aims to construct semi-/weak-/self- supervision signals from 3D data [8, 10, 11, 26, 28, 39, 60, 74, 77, 78, 86], so as to minimize the need for 3D annotations. In addition, some methods leverage 2D supervision to assist the model training for 3D scene understanding [20, 33, 45, 61, 68, 72].

However, the aforementioned methods all rely on training with domain-specific 3D or 2D data, limiting their zero-shot ability to understand new scenes that have never been seen during training. Instead, our framework seeks to straightforwardly harness the inherent zero-shot ability of SAM for segmenting 3D scenes, thereby eliminating the need for additional model training.

Zero-shot and open-set 3D scene understanding. The early studies on zero-shot 3D scene understanding are very limited [47, 80] and they still involve training with supervised 3D labels. In recent years, many 2D vision foundation models have shown their remarkable zero-shot recognition abilities [3, 32, 51, 56, 70, 87, 88]. This encourages the researchers to leverage them for 3D scene parsing. For example, [7, 50, 52, 66] all use CLIP [56] to extract pixel-wise features and align them with 3D space to realize language-guided segmentation of 3D scenes. [15, 16] utilize [69] to caption multi-view images for associating 3D and open-vocabulary concepts. Liu *et al*. [40] distill knowledge from [32, 87, 88] to apply self-supervised learning on outdoor scenes. UnScene3D [60] lifts DINO [3] to initialize 3D features for self-training.

Despite recent advancements in open-set methods, applying them directly to new 3D data remains challenging, since they still necessitate model tuning [60], 3D-2D distillation [7, 15, 16, 40, 52], or 3D/2D pretrained region proposal network [29, 43, 50, 66]. In contrast, our method just uses SAM [32] on RGB frames without requiring any of the aforementioned factors. This enables direct deployment for segmenting novel 3D scenes.

The Segment Anything model. The Segment Anything Model (SAM) [32] has brought a revolution for image segmentation. Trained on an astonishing SA-1B dataset, SAM can effectively segment unfamiliar images without further training. The distinguishing characteristic of SAM lies in its promptability, allowing it to accept various input prompts, which specify where or what is to be segmented in an image. Several recent works are striving to lift SAM into 3D visual tasks. Cen *et al*. [4] use SAM to segment target objects in NeRF [48] via one-shot manual prompting. Zhang *et al*. [84] define hand-crafted grid prompts on Bird's Eye View images and perform SAM for 3D object detection. SAM3D [81] employs automatic-SAM on individual 2D frames to generate 2D segmentation masks, which are then projected into 3D and gradually fused to derive the final 3D segmentation. Similarly, SAI3D [83] over-segments a 3D scene into superpoints, which are then progressively merged according to SAM masks. However, both SAM3D and SAI3D all assign frame-specific prompts that lack consistency, causing segmentation conflicts across frames, finally yielding subpar 3D segmentation results. Very recently, SAM has been applied to generate training labels [27] or graph annotations [22] for 3D segmentation, yet these methods still require training on domain-specific data.

Different from them, our key idea is to locate SAM prompts in 3D space so that the pixel prompts derived from different frames but projected by the same 3D prompt become harmonized in the 3D space, leading to the frame consistency of prompts and their masks, and bringing high-quality 3D segmentation.

3. Method

An overview of our framework is presented in Fig. 3, which is designed in a bottom-up manner.

3.1. 3D Prompt Proposal

3D prompt initialization. Given a point cloud $\mathbf{F} \in \mathbb{R}^{N\times 3}$ of a 3D scene with N points, we first employ furthest-point sampling (FPS) to sample M points as the

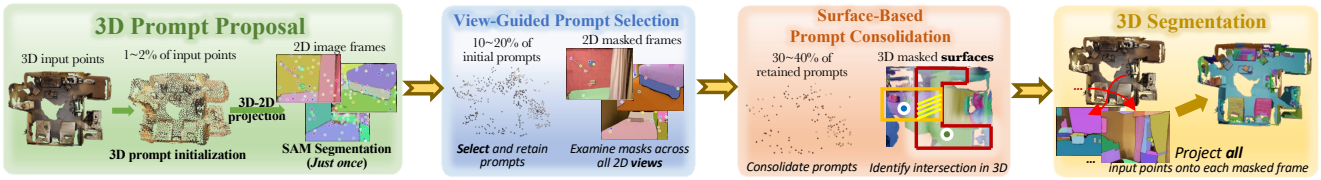

Figure 3. **An overview of our *SAMPro3D***, with a primary focus on ***"prompt"***. Given 3D scene point clouds with posed RGB-D frames, we locate SAM [32] prompts in input 3D scenes and project them onto 2D frames to obtain 2D segmentation masks. Later, the initial prompts and their masks are selected (Algorithm 1) and consolidated (Fig. 4), leveraging both multi-view and surface information. Finally, we project all input points onto each segmented frame to obtain the 3D segmentation result.

initial 3D prompt $\mathbf{P} \in \mathbb{R}^{M\times 3}$. FPS helps us to achieve a decent coverage of instances within a scene. For simplification, we use $\mathbf{f} \in \mathbb{R}^3$ and $\mathbf{p} \in \mathbb{R}^3$ to denote an individual input point and a single 3D prompt, respectively.

3D-2D projection. Following [52], we only consider the pinhole camera configuration here. In particular, given the camera intrinsic matrix I_i and world-to-camera extrinsic matrix E_i of a frame i, we calculate the corresponding pixel projection $\mathbf{x} = (u, v)$ of a point prompt $\mathbf{p}$ by: $\tilde{\mathbf{x}} = I_i \cdot E_i \cdot \tilde{\mathbf{p}}$, where $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{p}}$ are the homogeneous coordinates of $\mathbf{x}$ and $\mathbf{p}$, respectively. Similar to [52, 66], an occlusion test is performed using depth values, to ensure that the pixel $\mathbf{x}$ is only valid when its corresponding point $\mathbf{p}$ is visible in frame i.

SAM segmentation on image frames. SAM [32] is a promptable segmentation model that can accept various inputs such as pixel coordinates or bounding boxes and predict the segmentation area associated with each prompt. In our framework, we feed all pixel coordinates calculated before to prompt SAM and obtain the 2D segmentation masks on all frames. As depicted in Fig. 2 and described in Sec. 1, through locating prompts in 3D space, the pixel prompts originating from distinct frames but projected by the identical 3D prompt will be aligned in 3D space, bringing the frame consistency on pixel prompts and their SAM-predicted masks. Notably, this step is the *only* step that we need to perform SAM. In the later stages, our attention is directed toward selecting and consolidating initial 3D prompts along with their segmentation masks.

3.2. View-Guided Prompt Selection

After the prompt initialization, although consistent, multiple prompts may segment the same instance, causing redundancy in the segmentation. Besides, some prompts may generate inaccurate masks that hurt the performance. To handle this, we introduce a View-Guided Prompt Selection algorithm to select prompts.

As outlined in Algorithm 1, we examine SAM-predicted masks and accumulate the examinations across all views. First, in each view, if a 3D prompt $\mathbf{p}$ has a valid pixel projection $\mathbf{x}$, its counter $\mathbf{c}$ increments. Next, if its 2D mask does not have overlaps with other masks, we select it as necessary for segmenting an instance. If several 2D masks exhibit significant overlaps, we select the ones with the highest SAM confidence value as the most representative prompt. The score $\mathbf{s}$ of a selected prompt will be accumulated. After examining all views, we compute the probability of retaining a 3D prompt by $\boldsymbol{\theta} = \mathbf{s} \,/\, \mathbf{c}$, and retain the prompt when its probability exceeds a predefined threshold θ_{retain}.

Algorithm 1 - View-Guided Prompt Selection

$\mathbf{s} \leftarrow 0, \mathbf{c} \leftarrow 0$
while i is a frame **do** # start single-view examination
  if $\mathbf{p}$ has a valid pixel projection $\mathbf{x}$ in current frame i **then**
    $\mathbf{c} \leftarrow \mathbf{c} + 1$
  end if
  Perform prompt selection according to the information of their SAM-predicted masks
  if $\mathbf{x}$ is selected after this examination **then**
    $\mathbf{s} \leftarrow \mathbf{s} + 1$
  end if
  $i \leftarrow i + 1$ # go to the next frame
end while # finish examination on all frames
$\boldsymbol{\theta} = \mathbf{s} \,/\, \mathbf{c}$
if $\boldsymbol{\theta} > \theta_{retain}$ **then**
  Retain this 3D prompt $\mathbf{p}$
end if

This algorithm enables us to utilize multi-view information from all 2D views. It prioritizes high-quality prompts while maintaining prompt consistency, ultimately enhancing the 3D segmentation result. It is ablated in Sec. 4.4. More details are provided in the supplementary material.

3.3. Surface-Based Prompt Consolidation

We have observed that in some cases, a single retained prompt may only segment part of a 3D instance due to the limited visible field of 2D camera views. This issue is particularly prominent for large-sized instances that require multiple 2D views to be fully captured. As in Fig. 4 (a), several prompts segment different parts of the floor.

To address this, instead of solely using multi-view information, we further explicitly leverage 3D surface information and develop a Surface-Based Prompt Consolidation strategy (see Fig. 4 (b)). This strategy involves checking the 3D masked *surfaces* generated by different 3D prompts and identifying a certain intersection between them in 3D

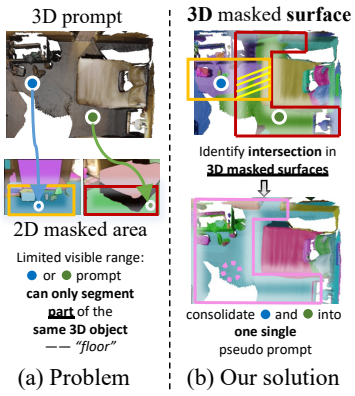

Figure 4. The illustration of the partial segmentation problem and our Surface-based Prompt Consolidation strategy.

space. In such cases, we consider these prompts as likely segmenting the same 3D instance and consolidate them into a single pseudo prompt. This process facilitates the integration of 3D information across prompts, leading to a more comprehensive segmentation of 3D instances.

The prompt consolidation performs stably well in our experiment. The key to consolidation lies in identifying the *intersection* between surfaces segmented by prompts across views. Since 2D sequences have small view changes (even when skipping 20 frames as in supplementary material), there is always an intersection between continuous masked surfaces that segment the same 3D instance.

3.4. 3D Scene Segmentation

After previous procedures, we have obtained the final set of 3D prompts and their 2D segmentation masks across frames. In addition, we have also ensured that each single prompt segment one 3D instance, allowing the *prompt ID* to naturally serve as the *instance ID*. With the ultimate goal of segmenting all points within the 3D scene, we continue by projecting *all* input points of the scene onto each segmented frame and compute their predictions using the following steps: For each individual input point $\mathbf{f}$ in the scene, if it is projected within the mask area segmented by a prompt p_k in frame i, we assign its prediction within that frame as the prompt ID k. We accumulate the predictions of $\mathbf{f}$ across all frames and assign its final prediction ID based on the prompt ID that has been assigned to it the most number of times. By repeating this for all input points, we can achieve a complete 3D segmentation of the input scene.

In our study, all points can successfully get a valid mask, as it is *easy to achieve*: i) Simply initializing adequate prompts ensures comprehensive scene coverage with their generated masks, yet using only a few prompts may cause segmentation absence (Fig. 7: 0.1%). ii) Even if some frames have pixels without masks, there are sufficient frames to provide masks. As in supplementary material, all points can get a mask although skipping 20 frames. iii) If a point fails to get a mask (not observed in our experiment), we can assign it the label that occurs most frequently among its neighboring labeled points.

4. Experiments

Setup. We conduct experiments on diverse 3D scene datasets, including ScanNet (v2) [13], ScanNet200 [59] and ScanNet++ [82] for indoor rooms, and KITTI-360 [37] for outdoor environments. ScanNet contains 1513 RGB-D indoor scans, with estimated camera parameters, surface reconstructions, textured meshes, and semantic annotations. ScanNet200 provides *more extensive annotations* for 200 categories based on ScanNet. ScanNet++ offers 280 indoor scenes with *more detailed segmentation masks* and high-resolution RGB images. ScanNet200 and ScanNet++ present the conspicuously challenging scenario for zero-shot scene segmentation. KITTI-360 is an *outdoor* scene dataset with 300 suburban scenes, comprising 320k images and 100k laser scans. We use their official validation split for evaluation. In the supplementary material, we also evaluate the robustness of our method on Matterport3D [5] dataset which exhibits *large view changes*.

To expedite processing, we resize each RGB image frame to a resolution of 240×320, which has proven to be sufficient for our method to generate high-quality 3D segmentation results. We employ the ViT-H SAM [32] model, which is the default public model for SAM. The entire framework is executed on a single NVIDIA A100 GPU. As our method is *class-agnostic*, we benchmark it against the class-agnostic baselines to ensure a fair comparison. The semantic labels are not considered in the evaluation, following [22, 60]. Additionally, we follow [22] to exclude the predicted instances in unannotated regions for all methods, facilitating a fairer comparison.

Methods in comparison. (1) **Open-set** methods, which do not learn from manually annotated 3D labels from a predefined set. These methods can be further split into two categories: i) **training-*dependent*** methods which require additional model tuning [22, 27, 60] on domain-specific data. ii) **training-*free*** methods without additional training, containing traditional unsupervised methods [18, 46] and SAM-based methods [81, 83]. As highlighted previously, our method directly applies SAM to RGB-frames, which also eliminates the need for further training.

Note that SAM3D [81] and Segment3D [27] additionally report the scores incorporating the result from Felzenszwalb's algorithm [18] to refine their original results. Here, to fairly compare the performance of the algorithms themselves, we present their results without post-processing.

(2) **Closed-set** method: we also compare our method with Mask3D [62], a state-of-the-art fully-supervised method trained on predefined annotations. We use its offi-

Model	*ScanNet* AP	AP_{50}	AP_{25}	*ScanNet200* AP	AP_{50}	AP_{25}	*ScanNet++* AP	AP_{50}	AP_{25}	*KITTI-360* AP	AP_{50}	AP_{25}
training-*dependent*												
UnScene3D [60] [CVPR'24]	15.9	32.2	58.5	-	-	-	-	-	-	-	-	-
Segment3D [27] [ECCV'24]	-	-	-	-	-	-	12.0	22.7	37.8	-	-	-
SAM-graph [22] [ECCV'24]	-	-	-	22.1	41.7	62.8	15.3	27.2	44.3	23.8	**37.2**	49.1
training-*free*												
HDBSCAN [46] [ICDMW'17]	1.6	5.5	32.1	2.9	8.2	33.1	4.3	10.6	32.3	9.3	18.9	39.6
Felzenszwalb [18] [IJCV'04]	5.3	12.6	36.9	4.8	9.8	27.5	8.8	16.9	36.1	-	-	-
SAM3D [81] [ICCVW'23]	6.3	17.9	47.3	12.1	38.6	54.1	3.0	7.9	22.3	4.6	10.6	26.0
SAI3D [83] [CVPR'24]	**30.8**	**50.5**	**70.6**	-	-	-	17.1	31.1	49.5	-	-	-
Ours	24.3	45.7	67.7	**26.3**	**47.2**	**68.6**	**20.3**	**35.6**	**53.2**	**24.3**	34.7	**52.8**

Table 1. **Quantitative** comparison with diverse **open-set** methods on ScanNet, ScanNet200, ScanNet++ and KITTI-360.

		Training Data	AP	AP_{50}	AP_{25}
ScanNet200	Mask3D	GT of ScanNet200	**53.3**	**71.9**	**81.6**
	Mask3D	GT of ScanNet	45.1	62.6	70.5
	Ours	-	26.3	47.2	68.6
ScanNet++	Mask3D	GT of ScanNet200	4.6	10.5	22.9
	Mask3D	GT of ScanNet	3.7	7.9	15.6
	Ours	-	**20.3**	**35.6**	**53.2**
KITTI-360	Mask3D	GT of ScanNet200	0.2	0.9	7.0
	Mask3D	GT of ScanNet	0.3	1.0	8.0
	Ours	-	**24.3**	**34.7**	**52.8**

Table 2. **Quantitative** comparison with **closed-set** method Mask3D [62]. While Mask3D outperforms ours on ScanNet200 when trained on ScanNet or ScanNet200, it cannot generalize well to ScanNet++ and KITTI-360.

cial pretrained models that are trained on ScanNet and ScanNet200 training sets. Additionally, Mask3D does not treat floor and wall as instances, resulting in the absence of these two labels in its results.

Quantitative metrics. We follow the standard of instance segmentation task defined in [13, 58], calculating mean Average Precision (mAP) at IoU thresholds of 50%, 25%, and averaged from 50% to 95% with a step size of 5% (denoted by $\mathbf{AP_{50}}$, $\mathbf{AP_{25}}$ and $\mathbf{AP}$, respectively).

4.1. Main Results

Quantitative Results (1) As in Tab. 1, when compared with **open-set** methods, our training-free approach achieves comparable or better performance than training-dependent/free methods. While the recent method SAI3D [CVPR'24] surpasses ours on ScanNet, our approach excels on ScanNet++ (4.5↑ $\mathbf{AP_{50}}$). To explain, ScanNet's annotations are generally coarse, whereas ScanNet++ offers detailed segmentation annotations for fine-grained instances. These detailed annotations hold more significance in evaluating *zero-shot* capabilities, demonstrating our method's proficiency in segmenting *fine-grained* instances, consistent with Fig. 5. Moreover, our training-*free* method outperforms the very recent training-*dependent* method SAM-graph [ECCV'24] on all datasets, especially on ScanNet200 (5.5↑ $\mathbf{AP_{50}}$) and ScanNet++ (8.4↑ $\mathbf{AP_{50}}$).

(2) Following [22], Tab. 2 reports the comparison with

Figure 5. **Qualitative comparison** of our method, SAM3D [81], Mask3D [62] and ScanNet200's annotations [59], across various scenes in ScanNet200, from holistic to focused view. Mask3D does not treat floor and wall as instances, resulting in the absence of these two labels in its results. Better view in zoom and color.

the **closed-set** method, Mask3D. When Mask3D is trained on ScanNet or ScanNet200 and then tested on ScanNet200, it outperforms our method. However, its results significantly drop and lag behind ours when tested on ScanNet++. ScanNet++ is collected by the laser scanner, which is different from ScanNet gained by RGB-D fusion[14]. This shows that Mask3D is sensitive to data acquisition schemes. Furthermore, when evaluated on the outdoor KITTI-360 which is distinct from indoor ScanNet200, Mask3D's performance becomes extremely poor. This comparison highlights the zero-shot power of our approach.

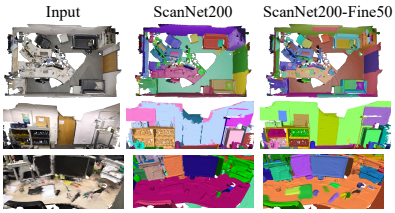

Figure 6. Samples from ScanNet200 and our **ScanNet200-Fine50**.

Instance Size	Normal			Small			Tiny		
	AP	AP_{50}	AP_{25}	AP	AP_{50}	AP_{25}	AP	AP_{50}	AP_{25}
Closed-set									
Mask3D [62]	**53.6**	**72.3**	**81.9**	9.3	17.4	30.1	4.2	13.5	16.8
Open-set									
SAM3D [81]	13.4	38.9	55.6	8.5	17.2	28.3	4.5	12.4	17.6
Ours	28.5	48.1	68.5	**17.8**	**30.3**	**40.7**	**14.3**	**25.6**	**35.2**

Table 3. Quantitative comparison with Mask3D (trained on ScanNet200) and SAM3D on our **ScanNet200-Fine50** test data across different instance sizes.

Qualitative results. Fig. 5 shows qualitative comparisons on ScanNet200 [59], across a variety of scenes (*e.g.*, bedroom, office, classroom) and instances, from the holistic to the focused perspective. Our method outperforms the open-set method SAM3D in terms of both segmentation quality and diversity. When compared to the closed-set approach Mask3D (*trained and evaluated both on ScanNet200*), our method achieves competitive or even superior segmentation accuracy and diversity. Moreover, compared to the extensive annotations of ScanNet200, our results are not only comparable in accuracy but also exhibit greater diversity in many cases. More results a *video demo* are provided in the supplementary material.

4.2. A Fine-grained Test Set – *ScanNet200-Fine50*

Nonetheless, we still cannot precisely evaluate our fine-grained predictions due to the lack of corresponding accurate GT annotations in available datasets. To remedy this defect, we select 50 scenes from ScanNet200 [59] validation set, and provide high-quality and fine-grained annotations, yielding *ScanNet200-Fine50* test set. A visual comparison between samples from the original ScanNet200 and our ScanNet200-Fine50 is provided in Fig. 6. More samples are shown in the supplementary material.

Similar to [27], we split ScanNet200-Fine50's annotations based on the mask size (*i.e.*, number of points) of each instance to emphasize the fine-grained performance, including tiny ($0 \sim 1k$), small ($1 \sim 2k$), normal ($2k \sim \infty$). As shown in Tab. 3, although the closed-set method Mask3D [62] (trained on ScanNet200) surpasses ours on normal-sized instances, our method excels on tiny (12.1 $\mathbf{AP_{50}}$ ↑) and small objects (12.9↑ $\mathbf{AP_{50}}$). In addition, our approach achieves superior results than the open-set method SAM3D [81] across all instance sizes. These results are also supported by Fig. 5, validating the zero-shot ability of our method on *highly fine-grained* segmentation. Our ScanNet200-Fine50 can serve as a supplementary test set to evaluate future zero-shot methods.

Method	SAM3D [81]	Mask3D [62]	Annotations [59]	**Ours**
mAcc	1.531±0.321	2.417±0.417	3.021±0.372	**3.035±0.412**
mDiv	1.552±0.491	2.308±0.420	2.909±0.409	**2.927±0.361**

Table 4. The quantitative results of **user study**. "mAcc" and "mDiv" denote mean scores of accuracy and diversity.

4.3. User Study

We further conduct a subjective user study following SAM [32], as a complementary assessment. We select a set of 20 reference results randomly, with most of them adjusted to a focused view for clearer discrimination. We invited 100 subjects through an online questionnaire. All participants had no prior experience with 3D scene understanding, and none of them had seen our results before. We present each subject with five images for each case, including qualitative results (on ScanNet200) of SAM3D, Mask3D, and ScanNet200's annotations, arranged side by side in random order, with an input image as a reference. During the evaluation, we instruct subjects to rank the four results based on two criteria: segmentation *accuracy* and *diversity*. The accuracy evaluates the clarity of the segmented boundaries, while diversity focuses on the extent of whether "segmented anything". The result ranked first by the subjects is assigned a score of 4, while the last-ranked result receives 1.

The mean scores with standard deviation for accuracy (mAcc) and diversity (mDiv) are presented in Tab. 4. The results are statistically significant, demonstrating that our method surpasses both SAM3D and Mask3D by a large margin. *Notably*, even when compared to ScanNet200's annotations, our method achieves slightly higher scores in both quality and diversity. This user study further confirms the efficacy and zero-shot ability of our method.

4.4. Ablation Studies and Analysis

Efficiency. Our pipeline exhibits good efficiency, with the majority of computational time and memory usage allocated to the inference process of SAM across the RGB frames. In terms of memory usage, a single GPU with approximately 8000MB is sufficient to run SAM [32] along with our entire pipeline.

Regarding computational time, Tab. 5 provides a breakdown of the time consumed by each step in our framework and compares ours with SAM3D [81]. Similar to SAM3D, our pipeline sequentially processes all frames, allowing room for speed improvement through parallel computation across frames. In SAM3D, each 2D mask must be projected into 3D masks, which are then iteratively merged based on k-nearest-neighbor search across adjacent frames until achieving the final 3D segmentation of the entire scene. This iterative process increases the time cost. As

Pro.	Sel.	Con.	Seg.	**Ours** (Total)	SAM3D [81]
10	2	1	2	**=15**	20

Table 5. **Running time** (in minutes) on the scene of ∼2,000 frames. "Pro.", "Sel.", "Con." and "Seg." respectively indicate prompt proposal, selection, consolidation, and 3D segmentation in our pipeline.

	Module Impact		**Initial Prompts**				θ_{retain}						**Selection**	
	w/o Sel.	w/o Con.	1%	1.5%	2%	5%	0.3	0.4	0.5	0.6	0.7	0.8	soft	top-k
AP	19.5	21.4	**26.3**	25.8	**26.3**	17.5	23.0	**26.3**	25.8	**26.3**	25.6	22.4	26.2	25.4
AP$_{50}$	40.8	41.2	**47.2**	**47.2**	46.9	39.8	43.0	**47.2**	**47.2**	46.9	46.2	42.3	47.0	47.1
AP$_{25}$	60.7	61.2	**68.6**	68.5	**68.6**	60.3	64.2	**68.6**	**68.6**	68.4	68.0	63.3	68.2	68.3

Table 6. The quantitative ablation studies on ScanNet200. "w/o Sel." and "w/o Con." respectively denote discarding prompt selection and consolidation. We also evaluate our method using different ratios (1%, 1.5%, 2%, 5%) of input points as our initial prompts. θ_{retain} is the threshold in prompt selection. "soft" and "top-k" are two voting schemes used during prompt selection.

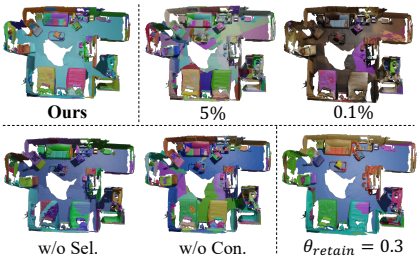

Figure 7. The qualitative ablation studies on ScanNet200.

for Mask3D [62], it needs ∼0.5s to perform inference on each scene, but requires ∼78 hours to train. Instead, our method does *not* require training.

Module impacts. As outlined in Tab. 6 and depicted in Fig. 7, removing View-Guided Prompt Selection results in a performance drop, and introduces inaccuracies in the examination of intersection areas during the prompt consolidation, thereby bringing fragmented segmentation. On the other hand, omitting Surface-Based Prompt Consolidation conspicuously causes fragmented segmentation.

The number of initial prompts. Our method performs stably well when an appropriate number of prompts is initialized (Tab. 6: 1%, 1.5%, 2% of input points). However, using too many initial prompts introduces more redundancy, which hurts the overall performance (Tab. 6: 5%), and also causes inaccuracies in the examination of intersection areas during the prompt consolidation, thereby bringing fragmented segmentation (Fig. 7: 5%). Using only a few initial prompts results in the absence of segmentation for many instances (Fig. 7: 0.1%).

θ_{retain} during prompt selection. During the View-Guided Prompt Selection (Algorithm 1), we define a threshold θ_{retain} to decide the probability of retaining a prompt. Our method demonstrates satisfactory performance across a flexible range of θ_{retain} (Tab. 6: $\theta_{retain} = 0.4, 0.5, 0.6, 0.7$). Furthermore, setting θ_{retain} to either a small or large value (Tab. 6: $\theta_{retain} = 0.3, 0.8$, Fig. 7: $\theta_{retain} = 0.3$) *weakens* the effect of prompt selection and causes performance drop, which conversely verifies the benefit of our selection algorithm.

The way of prompt selection. As for our View-Guided Prompt Selection (Algorithm 1), we also explored alternative selection methods. One approach involved using the actual scores obtained during the selection process on individual frames and averaging these scores across all frames. We compared the mean score with θ_{retain}, to determine whether a prompt should be retained. This method can be referred to as "soft" voting. Additionally, we tried another selection scheme where prompts are kept based on their frequency of being retained across all frames, referred to as the top-k voting scheme. As shown in Tab. 6, both the "soft" and top-k voting schemes yield competitive results. This indicates the stability and effectiveness of our method in utilizing different selection ways.

To summarize, the above ablations demonstrate that our method does *not* require a complex hyperparameter setup, highlighting its simplicity and effectiveness. The ablation results on our ScanNet200-Fine50 test set and more ablation studies are provided in the supplementary material.

5. Conclusion

We have proposed SAMPro3D for zero-shot segmentation of 3D scenes by utilizing SAM on 2D frames. The key idea is to locate SAM prompts in 3D to align pixel prompts across frames. Based on this key idea, we introduced a prompt selection algorithm and a prompt consolidation strategy to produce high-quality and comprehensive 3D segmentation. Our method does not need any additional training, preserving the zero-shot capability of SAM.

Acknowledgments. The work was supported in part by Guangdong Provincial Outstanding Youth Fund (No. 2023B1515020055), the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, the National Key R&D Program of China with grant No. 2018YFB1800800, Shenzhen Outstanding Talents Training Fund 202002, Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104, Key Area R&D Program of Guangdong Province (Grant No. 2018B030338001), the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), and Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055). It is also partly supported by NSFC-61931024, and Shenzhen Science and Technology Program No. JCYJ20220530143604010.

References

[1] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. 3

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 3

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1, 3

[4] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Chen Yang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, and Qi Tian. Segment anything in 3d with nerfs. In *NeurIPS*, 2023. 1, 3

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *3DV*, 2017. 3, 5

[6] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, 2020. 3

[7] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *CVPR*, 2023. 1, 3

[8] Yujin Chen, Matthias Nießner, and Angela Dai. 4dcontrast: Contrastive learning with dynamic correspondences for 3d scene understanding. In *ECCV*, 2022. 3

[9] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *CVPR*, 2023. 3

[10] Mingmei Cheng, Le Hui, Jin Xie, and Jian Yang. Sspc-net: Semi-supervised semantic 3d point cloud segmentation network. In *AAAI*, 2021. 3

[11] Julian Chibane, Francis Engelmann, Tuan Anh Tran, and Gerard Pons-Moll. Box2mask: Weakly supervised 3d semantic instance segmentation using bounding boxes. In *ECCV*, 2022. 3

[12] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In *CVPR*, 2019. 1, 3

[13] 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. 2, 3, 5, 6

[14] Angela Dai, Matthias Nießner, Michael Zollöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface re-integration. *TOG*, 2017. 6

[15] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. PLA: Language-Driven Open-Vocabulary 3D Scene Understanding. In *CVPR*, 2023. 1, 3

[16] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Lowis3d: Language-driven open-world instance-level 3d scene understanding. *PAMI*, 2024. 3

[17] Shichao Dong, Fayao Liu, and Guosheng Lin. Leveraging large-scale pretrained vision foundation models for label-efficient 3d point cloud segmentation. *arXiv preprint arXiv:2311.01989*, 2023. 1

[18] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004. 5, 6

[19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 3

[20] Kyle Genova, Xiaoqi Yin, Abhijit Kundu, Caroline Pantofaru, Forrester Cole, Avneesh Sud, Brian Brewington, Brian Shucker, and Thomas Funkhouser. Learning 3d semantic segmentation with only 2d image supervision. In *3DV*, 2021. 3

[21] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. 3

[22] Haoyu Guo, He Zhu, Sida Peng, Yuang Wang, Yujun Shen, Ruizhen Hu, and Xiaowei Zhou. Sam-guided graph cut for 3d instance segmentation. In *ECCV*, 2024. 1, 3, 5, 6

[23] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *CVPR*, 2021. 1

[24] Binh-Son Hua, Minh-Khoi Tran, and Sai-Kit Yeung. Pointwise Convolutional Neural Network. In *CVPR*, 2018. 3

[25] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *3DV*, 2016. 3

[26] Di Huang, Sida Peng, Tong He, Honghui Yang, Xiaowei Zhou, and Wanli Ouyang. Ponder: Point cloud pre-training via neural rendering. In *ICCV*, 2023. 3

[27] Rui Huang, Songyou Peng, Ayca Takmaz, Federico Tombari, Marc Pollefeys, Shiji Song, Gao Huang, and Francis Engelmann. Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels. In *ECCV*, 2024. 1, 3, 5, 6, 7

[28] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *ICCV*, 2021. 3

[29] Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. In *ECCV*, 2024. 1, 3

[30] Li Jiang, Hengshuang Zhao, Shu Liu, Xiaoyong Shen, Chi-Wing Fu, and Jiaya Jia. Hierarchical point-edge interaction network for point cloud semantic segmentation. In *ICCV*, 2019. 3

[31] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *CVPR*, 2020. 1, 3

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023. 1, 3, 4, 5, 7

[33] Abhijit Kundu, Xiaoqi Yin, Alireza Fathi, David Ross, Brian Brewington, Thomas Funkhouser, and Caroline Pantofaru. Virtual multi-view fusion for 3d semantic segmentation. In *ECCV*, 2020. 3

[34] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A composite dataset for multi-domain semantic segmentation. In *CVPR*, 2020. 1

[35] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. PointCNN: Convolution on X-transformed Points. In *NeurIPS*, 2018. 3

[36] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, Sai Bi, Hong-Xing Yu, Zexiang Xu, Kalyan Sunkavalli, Milos Hasan, Ravi Ramamoorthi, and Manmohan Chandraker. Openrooms: An open framework for photorealistic indoor scene datasets. In *CVPR*, 2021. 3

[37] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *PAMI*, 2022. 2, 5

[38] Yiqun Lin, Zizheng Yan, Haibin Huang, Dong Du, Ligang Liu, Shuguang Cui, and Xiaoguang Han. Fpconv: Learning local flattening for point convolution. In *CVPR*, 2020. 3

[39] Kangcheng Liu, Yuzhi Zhao, Qiang Nie, Zhi Gao, and Ben M Chen. Weakly supervised 3d scene segmentation with region-level boundary awareness and instance discrimination. In *ECCV*, 2022. 3

[40] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. In *NeurIPS*, 2023. 1, 3

[41] Ze Liu, Han Hu, Yue Cao, Zheng Zhang, and Xin Tong. A closer look at local aggregation operators in point cloud analysis. In *ECCV*, 2020. 3

[42] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *ICCV*, 2021. 3

[43] 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*, 2023. 3

[44] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. In *ICLR*, 2022. 3

[45] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *ICRA*, 2017. 3

[46] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *ICDMW*, 2017. 5, 6

[47] Björn Michele, Alexandre Boulch, Gilles Puy, Maxime Bucher, and Renaud Marlet. Generative zero-shot learning for semantic segmentation of 3D point cloud. In *3DV*, 2021. 1, 3

[48] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

[49] Ishan Misra, Rohit Girdhar, and Armand Joulin. An End-to-End Transformer Model for 3D Object Detection. In *ICCV*, 2021. 3

[50] Phuc D. A. Nguyen, Tuan Duc Ngo, Chuang Gan, Evangelos Kalogerakis, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *CVPR*, 2024. 1, 3

[51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[52] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D Scene Understanding with Open Vocabularies. In *CVPR*, 2023. 1, 3, 4

[53] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *CVPR*, 2017. 3

[54] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *NeurIPS*, 2017. 3

[55] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *ICCV*, 2019. 3

[56] 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. 1, 3

[57] Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197*, 2023. 2

[58] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 6

[59] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *ECCV*, 2022. 2, 5, 6, 7

[60] David Rozenberszki, Or Litany, and Angela Dai. Unscene3d: Unsupervised 3d instance segmentation for indoor scenes. In *CVPR*, 2024. 3, 5, 6

[61] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *CVPR*, 2022. 3

[62] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In *ICRA*, 2023. 1, 3, 5, 6, 7, 8

[63] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*, 2019. 3

[64] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 3

[65] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 3

[66] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *NeurIPS*, 2023. 1, 3, 4

[67] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, 2019. 1

[68] Vibhav Vineet, Ondrej Miksik, Morten Lidegaard, Matthias Nießner, Stuart Golodetz, Victor A Prisacariu, Olaf Kähler, David W Murray, Shahram Izadi, Patrick Pérez, et al. Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In *ICRA*, 2015. 3

[69] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, 2022. 3

[70] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. In *ICCV*, 2023. 3

[71] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph.*, 2019. 3

[72] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *CoRL*, 2022. 3

[73] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *CVPR*, 2019. 3

[74] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *ECCV*, 2020. 1, 3

[75] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *CVPR*, 2021. 1, 3

[76] Mutian Xu, Pei Chen, Haolin Liu, and Xiaoguang Han. Toscene: A large-scale dataset for understanding 3d tabletop scenes. In *ECCV*, 2022. 3

[77] Mingye Xu, Mutian Xu, Tong He, Wanli Ouyang, Yali Wang, Xiaoguang Han, and Yu Qiao. Mm-3dscene: 3d scene understanding by customizing masked modeling with informative-preserved reconstruction and self-distilled consistency. In *CVPR*, 2023. 3

[78] Xun Xu and Gim Hee Lee. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In *CVPR*, 2020. 3

[79] Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. SpiderCNN: Deep Learning on Point Sets with Parameterized Convolutional Filters. In *ECCV*, 2018. 3

[80] Yuwei Yang, Munawar Hayat, Zhao Jin, Hongyuan Zhu, and Yinjie Lei. Zero-shot point cloud segmentation by semantic-visual aware synthesis. In *ICCV*, 2023. 3

[81] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. In *ICCVW*, 2023. 1, 3, 5, 6, 7, 8

[82] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 2, 5

[83] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *CVPR*, 2024. 3, 5, 6

[84] Dingyuan Zhang, Dingkang Liang, Hongcheng Yang, Zhikang Zou, Xiaoqing Ye, Zhe Liu, and Xiang Bai. Sam3d: Zero-shot 3d object detection via segment anything model. *arXiv preprint arXiv:2306.02245*, 2023. 3

[85] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021. 1, 3

[86] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *CVPR*, 2020. 3

[87] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized decoding for pixel, image, and language. In *CVPR*, 2023. 3

[88] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *NeurIPS*, 2023. 3