Segment3D: Learning Fine-Grained Class-Agnostic 3D Segmentation without Manual Labels

Rui Huang[1] Songyou Peng[2] Ayça Takmaz[2] Federico Tombari[3] Marc Pollefeys[2,4]
Shiji Song[1] Gao Huang[1] Francis Engelmann[2,3]
[1]Tsinghua University [2]ETH Zurich [3]Google [4]Microsoft

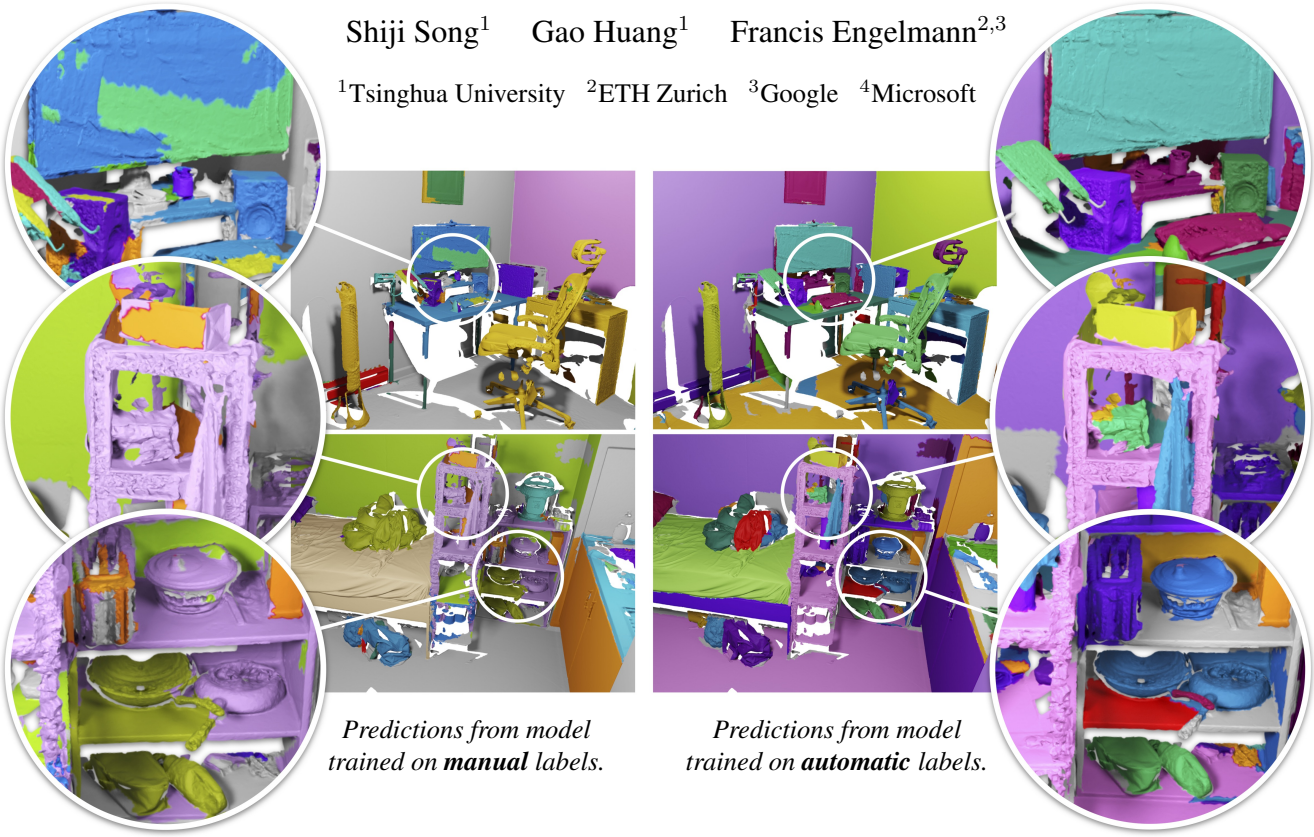

Figure 1. **Fine-Grained Class-Agnostic 3D Point Cloud Segmentation.** Segment3D predicts highly accurate segmentation masks *(right)*, improves over state-of-the-art 3D segmentation methods (e.g., Mask3D [40], *left*), and does not require manually labeled 3D training data. This is achieved through the automatic generation of high-quality training masks using foundation models for image segmentation [25].

Abstract

In this work, we aim to build a method for class-agnostic 3D scene segmentation. Current 3D scene segmentation methods largely rely on manually annotated 3D training datasets. This manual annotation process is labor intensive, often does not include fine details, and the models trained on it typically do not generalize well to new domains. Instead, we explore the use of image segmentation foundation models to automatically generate training labels for 3D segmentation. The resulting model, Segment3D, produces high-quality 3D segmentation masks, improves over existing 3D segmentation models (especially on fine-grained masks), and new training data can be easily added which further boosts the segmentation performance - all without the need for manual data labeling.

1. Introduction

In this work, we propose Segment3D, a method for fine-grained class-agnostic 3D segmentation. Understanding the complexity of the 3D world around us and dividing it into coherent segments aligned with both the scene geometry and its semantics is a cornerstone of machine perception. This ability to accurately segment and interpret 3D scenes is fundamental for intelligent non-human assistants, autonomous robots and AR/VR devices that help vision-impaired people navigate and engage with unknown spaces.

Current methods for 3D indoor-scene understanding mostly focus on semantic [31, 35, 36, 41, 45] and instance segmentation [8, 15, 27, 28, 40, 46]. These approaches, while effective in benchmarking, have limitations stemming from their training. Primarily, they depend on extensive manually labeled 3D training sets that are both time-consuming and

challenging to annotate. Additionally, their performance often deteriorates when applied to scenarios beyond their training data, limiting their effectiveness in diverse, real-world scenarios. This becomes particularly apparent under the recently emerging task of *open-vocabulary* 3D scene understanding [24, 29, 34, 44] that aims to segment arbitrary user queries, which naturally go beyond the pre-defined set of training-set classes. Concurrently, the recent surge in foundation models, particularly 2D vision-language models [22, 25, 37], demonstrates remarkable potential. Trained on internet-scale data, these models exhibit an extraordinary ability to generalize, even in a zero-shot setting, to new and different input distributions. However, their application has been predominantly confined to 2D data. For instance, SAM [25] has shown impressive results in 2D image segmentation, but its applicability to 3D scene understanding remains mostly unexplored. All these factors give rise to the interesting and significant research question:

How to leverage foundation models for class-agnostic 3D scene segmentation without manually labeled 3D data?

To address this question, we introduce Segment3D, a novel approach that harnesses the strengths of 2D foundation models to achieve class-agnostic, fine-grained 3D segmentation. Segment3D employs a two-stage training approach that requires no hand-annotated labels for supervision. First, 2D masks are generated using SAM [25]. These masks are projected onto partial RGB-D point clouds and serve as supervision signal for pre-training our class-agnostic 3D segmentation model. This pre-training stage lays the groundwork for understanding the 3D structure from 2D annotations. However, as our ultimate objective is the segmentation of full 3D scenes, we must bridge the domain gap between partial point clouds and the more comprehensive 3D point clouds obtained from 3D scanners or reconstruction techniques [13, 21]. To this end, in the second stage, we fine-tune the model on full 3D point clouds in a self-supervised manner, utilizing high-confidence mask predictions from the pre-trained model as training signal.

In experiments, we show strong performance in class-agnostic segmentation on ScanNet++ [50] and further show the use of Segment3D for improving open-vocabulary 3D instance segmentation [44].

Overall, the contributions of this paper are as follows:

- We introduce Segment3D, a novel approach and training strategy for fine-grained class-agnostic 3D point cloud segmentation without manually annotated labels.
- Improved segmentation performance compared to a wide range of baselines, including fully supervised methods trained on carefully annotated datasets.
- We show that utilizing 2D foundation models to automatically generate high-quality training masks is a viable alternative to costly manual labeling.

2. Related Work

3D Instance Segmentation. In this domain, methods are mainly divided into three groups: proposal-based [48, 51], grouping-based [8, 23, 27, 46], and transformer-based [28, 40]. A representative work of proposal-based methods is GSPN [51], which generates high-quality 3D proposals and subsequently refines them through a region-based PointNet [35]. Grouping-based methods, such as PointGroup [23], predict per-point geometric offsets, then cluster points into instances, and finally output scores for individual instances. Mask3D [40] introduces a transformer-based framework for 3D instance segmentation, achieving state-of-the-art performance. In Mask3D, each object instance is represented as an instance query. Through the transformer decoders, these instance queries are refined by progressively attending to point cloud features across multiple scales. Despite these advancements, most studies have concentrated on fully supervised models reliant on ground-truth semantic labels. However, in the context of open-world 3D scene understanding, the importance of semantic classification for closed-set categories has diminished. Instead, the generalization to effectively segment a diverse range of objects has become particularly crucial.

Foundation Models. Foundation models, particularly those that are multimodal [22, 37], have revolutionized the field of AI by leveraging extensive image-text pre-training. These models can derive rich image representations guided by natural language descriptions, enabling them to adapt seamlessly to a variety of downstream tasks in a zero-shot manner. This adaptability extends to areas like object detection [18], semantic segmentation [38], and image manipulation [33], demonstrating a remarkable level of versatility. Another line of work [6, 32], based on self-supervised learning, employs image training and yields high-performance features directly applicable as inputs for straightforward linear classifiers. Segment Anything Model (SAM) [25] has recently advanced the performance of foundation models for image segmentation. SAM has undergone training on a diverse, high-quality dataset comprising more than 1 billion masks. This training equips SAM with the ability to zero-shot generalize to novel object types and images, surpassing the scope of its observations during the training process. Additionally, SAM can generate high-quality, fine-grained masks. Concurrent efforts alongside SAM [47, 53, 55, 56] also present opportunities for leveraging the foundation models across various downstream tasks. Our work leverages the power of SAM for pre-training a 3D network, which is later also used for generating supervision signals for fine-tuning on scene-level point clouds.

Open-Vocabulary 3D Scene Understanding. Recently, there has been an increased interest in 3D open-vocabulary scene understanding. This new field leverages the zero-

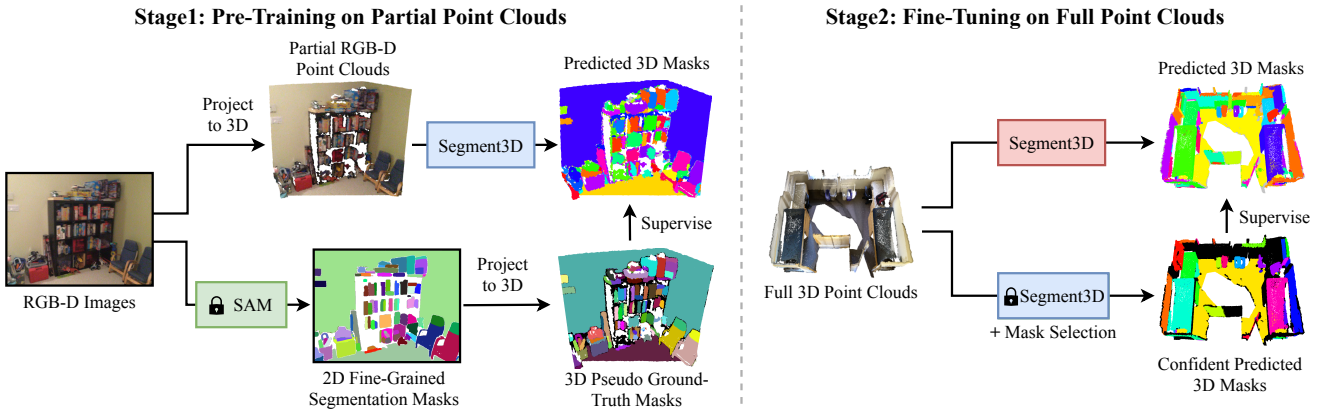

Figure 2. **Method Overview.** Training Segment3D involves two stages: The first stage *(left)* relies on largely available RGB-D image sequences and SAM, a pre-trained foundation model for 2D image segmentation [25]. Segment3D is pre-trained on partial RGB-D point clouds and supervised with pseudo ground-truth masks from SAM projected to 3D. Due to the domain gap between partial and full point clouds, in the second stage *(right)*, Segment3D is fine-tuned with confident masks predicted by the pre-trained Segment3D.

shot recognition capabilities of 2D vision-language models (VLM) [22, 37], enabling a more comprehensive understanding of diverse and previously unseen 3D environments [7, 14, 19, 20, 24, 26, 29, 34, 44, 52, 54]. PLA [14] aligns point cloud features with captions extracted from multi-view images of a scene to enable open-vocabulary recognition. CLIP2 [52] assembles triplets consisting of language descriptions, 2D images, and 3D point clouds. It utilizes a contrastive learning objective including semantic-level text-3D correlation and instance-level image-3D correlation. OpenScene [34] distills per-pixel image features to 3D point clouds, generating point-wise scene representations co-embedded with text and image pixels in CLIP feature space. However, it mainly focuses on semantic segmentation and exhibits a limited understanding of object instances. To this end, OpenMask3D [44] predicts class-agnostic 3D instance masks and aggregates per-mask features via multi-view fusion of CLIP-based image embeddings. Although they have achieved impressive results in open-vocabulary tasks, a highly versatile segmentation model is required to accomplish this. Our method contributes to this area by providing a variety of class-agnostic masks. These masks can serve as foundational inputs for models like OpenMask3D, enhancing their ability to perform instance-related tasks.

3. Method

We are interested in a method that can segment any object in a given 3D scene. Hence, depending on existing 3D training datasets cannot accomplish this goal, as their mask annotations are limited to a predefined set of object classes [1, 2, 12, 42] and models trained on these datasets might not generalize well to novel classes. Instead, the key idea is to automatically generate class-agnostic mask annotations using pre-trained foundation models [4]. The recently proposed SAM [25] is a foundation model for 2D image segmentation. To employ SAM for our 3D segmentation model we formulate a two-stage training approach as illustrated in Fig. 2. We first pre-train our class-agnostic 3D segmentation model with automatically generated 2D masks from SAM which are projected to *partial* RGB-D point clouds (Sec. 3.1). Since the final goal is to segment 3D scenes, we need to bridge the inevitable domain gap between partial RGB-D point clouds and full 3D point clouds from 3D scanners or reconstruction methods. We therefore fine-tune our model on *full* 3D point clouds using high-confident mask predictions from our pre-trained model (Sec. 3.2).

3.1. Stage 1: Pre-Training on RGB-D Point Clouds

In contrast to the relatively scarce availability of 3D data, there is an abundance of 2D data, particularly RGB-D images, which are readily accessible. For example, ScanNet [12] comprises merely 1513 3D scans, in contrast to the substantially larger collection of 2.5 million RGB-D images. Leveraging the automatically generated masks from SAM together with the abundant 2D data, we pre-train our 3D segmentation model on partial RGB-D point clouds.

Data Preparation. Starting from a collection of RGB-D frames, we create partial 3D point clouds and their corresponding pseudo ground-truth 3D masks. Note that the labels can be automatically obtained without manual effort.

For each frame in a large RGB-D dataset, we first transform the 2D depth map to a partial 3D point cloud. To do so, we need to know the intrinsic matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$ and extrinsic matrix $\mathbf{T} = \begin{bmatrix}\mathbf{R} & \mathbf{t}\end{bmatrix} \in \mathbb{R}^{3\times4}$. For each pixel $\mathbf{p} = (u, v)$, we can transform it with its depth value $D_\mathbf{p}$ into a 3D point $\mathbf{P}$ in world coordinates as follows:

$$\mathbf{P} = \mathbf{R}^\top \cdot (D_\mathbf{p} \cdot \mathbf{K}^{-1} \cdot \tilde{\mathbf{p}}) - \mathbf{R}^\top \mathbf{t} \quad (1)$$

where $\tilde{\mathbf{p}}$ is the homogeneous coordinate of $\mathbf{p}$. By applying the process in Eq. (1) to all pixels in the depth map D and associating their per-pixel RGB value, we obtain the input partial 3D point cloud. Now, to obtain the pseudo ground-truth 3D segmentation masks for this point cloud, we prompt SAM [25] with a regular grid on the RGB image and acquire on average ~ 50 high-quality 2D segmentation masks per frame. Since we know the one-to-one mapping between 2D pixel and 3D points from Eq. (1), we directly obtain the per-point 3D mask labels.

Model Architecture. We use a model inspired by Mask3D [40] to train a class-agnostic 3D segmentation model. The model is comprised of a sparse convolutional backbone derived from MinkowskiUNet [11] and a transformer decoder, as in MaskFormer [9, 10]. We adopt a set of queries to represent the masks, each of which is initialized with a positional embedding. Specifically, we select query positions with furthest point sampling (FPS) and use their Fourier positional encodings as the query embeddings. Leveraging the transformer decoder, all the mask queries are refined by progressively attending to point cloud features across multiple scales in parallel. Each mask query is subsequently decoded into both a mask feature and a binary label to predict whether the given query corresponds to a valid object or not. By computing cosine similarity scores between a mask feature and all point features within the point cloud, a heatmap is generated over the point cloud. This heatmap is input to a sigmoid function,and thresholded at 0.5, resulting in the final binary mask.

Training with SAM Generated Masks. We supervise the model with two losses: the per-point level supervision loss $\mathcal{L}_{\text{mask}}$ and a per-query level supervision loss $\mathcal{L}_{\text{obj}}$. The loss $\mathcal{L}_{\text{mask}}$ enables learning a foreground-background segmentation for each mask, and is composed of a dice loss $\mathcal{L}_{\text{dice}}$ [30] and a binary cross-entropy loss $\mathcal{L}_{\text{ce}}$ for each point. The $\mathcal{L}_{\text{obj}}$ is a binary classification loss that indicates whether a query represents a valid "object" or "no object". This mechanism allows for the prediction of a variable number of masks, depending on the underlying scene content and geometry. Following prior work [5, 9, 40], we first adopt bipartite graph matching to establish correspondences between the set of predicted masks and the set of target masks originating from SAM as described before. If the predicted instance finds a matching target mask, then we assign it an "object" label; conversely, if there is no match, we assign "no object". In summary, we optimize the following losses:

$$\mathcal{L} = \mathcal{L}_{\text{mask}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} \tag{2}$$

with

$$\mathcal{L}_{\text{mask}} = \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} \tag{3}$$

where λ_* are hyperparameters that balance the contribution of each component in the loss. The binary classification loss $\mathcal{L}_{\text{obj}}$ is applied to all queries, while the mask loss $\mathcal{L}_{\text{mask}}$ is specifically applied to masks labeled as "object".

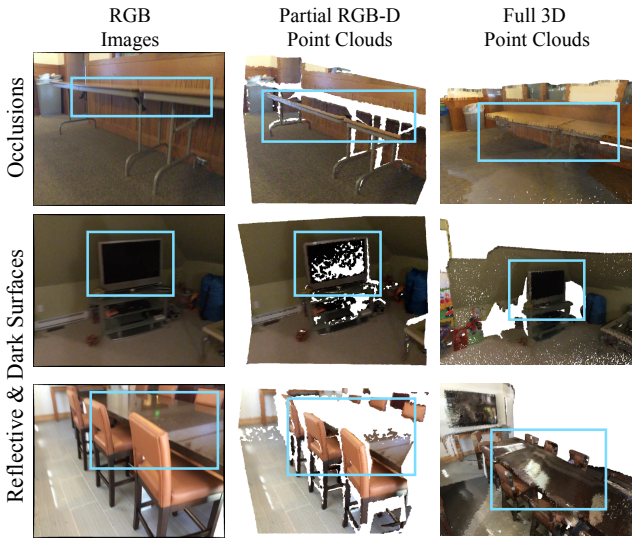

Figure 3. **Partial vs. Full Point Clouds.** We show the difference between partial *(center)* and full *(right)* point clouds from ScanNet [12]. Full point clouds are more complete and exhibit fewer occlusions due to reconstruction over multiple viewpoints. Due to this domain gap, fine-tuning on full 3D point clouds is necessary.

3.2. Stage 2: Self-Supervised Scene Fine-Tuning

After Stage 1, we obtain a class-agnostic 3D segmentation model by pre-training solely on RGB-D images and automatically generated labels from SAM. However, a fundamental domain gap persists between partial point clouds derived from RGB-D images and full point clouds acquired through a 3D scanner or reconstruction methods [13, 21] (See Fig. 3). This gap exists mostly because of object occlusions, but also due to challenges of depth cameras to capture dark or reflective surfaces from a single viewpoint. Hence, depending solely on RGB-D frames for training a 3D segmentation model intended for full 3D scans proves inadequate. Therefore, we propose to further fine-tune our model on scene-level full 3D point clouds. The key idea to obtain 3D mask annotations for training on full point clouds is to use selected, high-confidence masks generated by the pre-trained model itself. Note that this approach does not require any manual labels on full 3D scenes and proves essential for the performance of Segment3D (see Tab. 3).

Confidence-Based Mask Generation. Next, we outline the process of generating the supervision signal for the fine-tuning stage. The pre-trained model processes point clouds independently of their nature, be it partial or full. Therefore, when presented with a full 3D point cloud, the pre-trained 3D model produces a set of masks, each with a binary classification (valid or not) and a heatmap over all points, just as in Sec. 3.1. To assess the quality of the predicted masks, we

compute a confidence score based on the confidence map $\sigma(\mathbf{h})$, where $\mathbf{h}$ is the predicted heatmap and σ is the sigmoid function. We then compute the average confidence of those points for which $\sigma(\mathbf{h}) > 0.5$ as the confidence score of the predicted mask, denoted as c_{mask}. We also consider the classification as valid object and use the probability from the binary classification assigned to the "object" category as the confidence score and denoted it as c_{obj}. The final confidence score for each predicted mask is then the product of the two scores, $c = c_{\text{mask}} \cdot c_{\text{obj}}$. For fine-tuning our 3D segmentation model in Stage 2, we select the most confidently predicted masks above a threshold τ_c.

Training with the High-Confidence Generated Masks. For fine-tuning, we follow the same procedure as before and use $\mathcal{L}_{\text{mask}}$ as defined in Eq. 3. In contrast to the pre-training stage, the binary classification loss $\mathcal{L}_{\text{obj}}$, responsible for categorizing queries into valid or invalid, is omitted. Since we only select masks with high confidence for supervision, it can happen that some objects in the scene have no assigned ground truth mask. In such instances, deeming a correctly predicted mask for those objects as invalid would be detrimental. Table 3 illustrates the efficacy of the self-supervised fine-tuning process in comparison to pre-training alone.

Implementation Details. The feature backbone of Segment3D is a Minkowski Res16UNet34C [11]. We perform standard data augmentations, including horizontal flipping, random rotations, elastic distortion and random scaling. In addition, we use color augmentations including jittering, brightness and contrast augmentation. For Stage 1, we use AdamW optimizer and a one-cycle learning rate schedule with a peak learning rate of 2×10^{-4}. The model is trained for 20 epochs with a batch size of 16 partial RGB-D point clouds. A training on 2 cm voxelization takes approximately 60 hours with 2 RTX3090 GPUs. For Stage 2, the initial learning rate is set to 1×10^{-4}. We train the model for 50 epochs with a batch size of 8 full 3D point clouds. Training takes $\sim$ 10 hours with 2 cm voxels on 4 A100 GPUs.

4. Experiments

We firstly compare with fully-supervised and traditional geometric segmentation baselines in a class-agnostic segmentation setting (Sec. 4.1). We then provide detailed analysis to understand the importance of fine-tuning, the effect of training on more data and show the advantage of our method in segmenting small objects (Sec. 4.2). We also demonstrate its potential application for the task of open-set 3D instance segmentation as recently proposed in OpenMask3D [44] (Sec. 4.4). Finally, we show qualitative results of our 3D segmentation method (Sec. 4.3).

4.1. Comparing with State-of-the-Art Methods

Datasets. We run experiments on four popular datasets including ScanNet [12], its extension ScanNet200 [39], the Replica [42] dataset, and the newly released ScanNet++ [50]. For training our model, we employ ScanNet [12, 39], which is collected through a lightweight RGB-D scanning process. ScanNet comprises 1513 indoor scenes, encompassing $\sim 2 \cdot 10^6$ views, along with 3D camera poses, surface reconstruction and instance-level semantic mask annotations. For Stage 1, we sample every 25$^{\text{th}}$ frame of the RGB-D sequences ($\sim$ 1 FPS) and obtain approximately $76 \cdot 10^3$ training frames. For Stage 2, we use the $\sim 1.2 \cdot 10^3$ reconstructed 3D scans of indoor spaces as full point clouds. For evaluation, we use ScanNet++ [50] which comes with high-fidelity 3D mask annotations including smaller objects which are not well annotated in ScanNet. It includes high-resolution 3D scans acquired by a Faro laser scanner and high-resolution color images from an iPhone. The laser scans are captured at sub-millimeter precision and annotated comprehensively, covering objects of varying sizes.

Methods in Comparison. We compare with a wide range of prior art methods from different categories. Mask3D [40] is a fully-supervised transformer-based method trained on manually annotated 3D segmentation masks. Transformer-based methods [41, 43] currently define the SOTA for 3D instance segmentation. Segment3D has the same backbone as Mask3D but instead of training on manually annotated 3D masks, it learns from automatically generated 2D (pre-training) and 3D masks (fine-tuning).

Felzenszwalb *et al.* [17] proposed a graph-based method for segmentation which does not require any training data and operates directly on the 3D geometry.

A straightforward baseline to Segment3D is a model that directly projects 2D SAM masks onto the 3D point cloud and merges overlapping masks from different viewpoints. Such a baseline is already implemented in SAM3D [49], which predicts SAM masks in RGB images, then projects them onto partial point clouds and employs a bottom-up merging strategy to obtain 3D masks on the full point cloud. Similarly to ours, SAM3D does not rely on manual labels and can be seen as a non-learned variation of Segment3D which is essential to distill the improvement due to training.

Metrics. We evaluate all methods on the validation set of ScanNet++ and report average precision (AP) scores at IoU thresholds of 25%, 50%, and averaged over the range [0.5:0.95:05] between predicted and ground truth masks. Consistent with common practices in the field [41, 43, 46], we also report scores after post-processing. This involves smoothing the predicted masks through graph-based over-segmentation [17], and splitting distant parts of the same mask via connected component clustering DBScan++ [16].

Model	Ground Truth Labels	*without post-processing*			*with post-processing*		
		AP	**AP$_{50}$**	**AP$_{25}$**	**AP**	**AP$_{50}$**	**AP$_{25}$**
SAM3D [49]	✗	3.9	9.3	22.1	8.4	16.1	30.0
Felzenszwalb *et al.* [17]	✗	5.8	11.6	27.2	–	–	–
Mask3D [40]	ScanNet200 [39]	8.7	15.5	27.2	14.3	21.3	29.9
Mask3D [40]	ScanNet [12]	9.4	16.8	28.7	15.4	22.7	31.6
Segment3D (**Ours**)	✗	**12.0** (+27.7%)	**22.7** (+35.1%)	**37.8** (+31.7%)	**19.0** (+23.4%)	**29.7** (+30.8%)	**41.6** (+31.6%)

Table 1. **3D Segmentation Scores on ScanNet++ Val Set.** Evaluation metric is average precision (AP). Similar to [17, 49], Segment3D does not require manually annotated training labels. We report scores with and without post-processing (more details in Sec. 4.1).

Results. Scores are reported in Table 1. Segment3D outperforms all previous methods by at least $+\mathbf{23.4}\%$ AP and up to $+\mathbf{35.1}\%$ AP$_{50}$. Notably, we achieve such improvements *without ground-truth mask annotations* as used by Mask3D. In general, the performance of Mask3D (and other fully supervised methods) depends on the quality of the annotated training dataset; often the manual annotation of small objects (pens, cell-phones) or other fine-details is challenging. Instead, Segment3D relies on automatically generated high-quality masks from SAM which can capture fine-grained details without human annotation effort. Table 2 highlights that Segment3D excels particularly in predicting the more challenging small object masks.

Finally, our approach significantly outperforms SAM3D despite both methods relying on the same 2D masks from SAM. Even without fine-tuning Segment3D already outperforms SAM3D (see Table 3). This shows that a straightforward merging of 2D projections into 3D is insufficient, and that reasoning in 3D is essential.

Note the comparable performance of Mask3D trained on ScanNet and ScanNet200, which we attribute to similar mask annotations in both datasets. They differ only in the labeled classes (20 *vs.* 200), while most categories of ScanNet200 are labeled as "other" in ScanNet. For both datasets, we utilize all available masks independent of their semantic class to maximize the number of training labels.

Mask Size		**Mask3D** [40]	**Segment3D** (Ours)
Large	AP$_{50}$	30.6	30.4 (-0.65%)
	AP$_{25}$	47.7	47.0 (-1.47%)
Medium	AP$_{50}$	25.8	29.2 (+13.2%)
	AP$_{25}$	41.8	49.5 (+18.4%)
Small	AP$_{50}$	4.5	15.9 (+253%)
	AP$_{25}$	11.3	29.6 (+162%)

Table 2. **Segmentation Scores on Different Mask Sizes.** Evaluation metric is average precision (AP) on ScanNet++ validation set. Segmented3D improves over Mask3D, especially on small and medium-sized object masks. Details are in Sec. 4.2.

4.2. Analysis Experiments

Performance on Different Mask Sizes. We proceed with an analysis of the performance of our Segment3D and Mask3D (the best performing baseline) on object masks of various sizes. We categorize the size of masks based on the number of points they contain (small: [0, 2k], medium: [2k, 15k] large: [15k, ∞]) and exclude the masks of the floor, ceiling and walls. The results are reported in Table 2. Our method yields significantly improved segmentation results on small and medium-sized masks, and performs on-par with Mask3D on large masks. This confirms our intuition that Mask3D performs poorly on small-sized object masks as those are typically harder to manually annotate. In contrast, Segment3D utilizes masks from SAM as supervision, which capture fine-grained scene details. In summary, our model, trained on automatically generated masks by a foundation model for image segmentation, surpasses a model trained on manually labeled datasets. This showcases the usefulness of foundation models and raises the question if manually labeled large-scale 3D datasets are necessary for training 3D point cloud segmentation models.

The Effect of Two-Stage Training. Next, we compare the performance of Segment3D pre-trained solely on partial RGB-D point clouds (Stage 1) and with additional fine-tuning on full point clouds (Stage 2). Scores are reported in Table 3. The additional fine-tuning stage almost doubles the segmentation performance of our model on the most challenging AP metric. By training with the predicted high-confidence masks, Segment3D effectively reduces the inherent domain gap between the partial point clouds derived from RGB-D images and full 3D point clouds.

Training Stages	**AP**	**AP$_{50}$**	**AP$_{25}$**
Pre-Training (Stage 1)	7.4	15.2	31.2
+ Fine-Tuning (Stage 2)	**12.0** (+62%)	**22.7** (+49%)	**37.8** (+21%)

Table 3. **Effect of Two-Stage Training.** Fine-tuning on full point clouds supervised by confident predicted 3D masks significantly surpasses pre-training on projected 2D SAM masks alone.

Figure 4. **Qualitative Results on ScanNet++ Val Set.** From top to bottom, we show the colored input 3D scenes, the segmentation masks predicted by SAM3D [49], Mask3D [40], our Segment3D and the ground truth 3D mask annotations.

"a black eraser" *"kettle handle"* *"copier control screen"*

Figure 5. **Qualitative Results of Adapted OpenMask3D [44].** Given a text prompt *(bottom)*, OpenMask3D finds the corresponding masks ■ in a given 3D scene *(top)*. We show the 3D scene reconstruction and an RGB image for better visualization *(top left corner)*. The original OpenMask3D based on Mask3D, is unable to recognize the above queries, whether it is a relatively small object or a fine-grained affordance part of an object. In contrast, our Segment3D adaptation performs well in these cases.

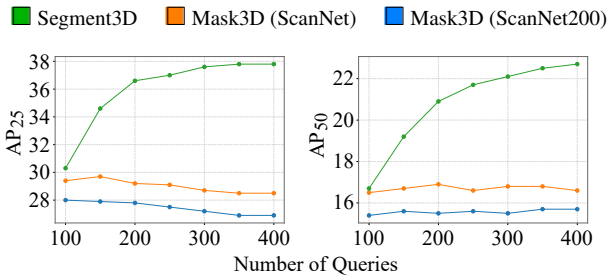

Figure 6. **Performance with Varying Number of Queries.** Metric is average precision (AP) with IoU threshold of 25% and 50%.

Number of Queries. Since our model and Mask3D share the same transformer-based architecture, we analyse the influence of the number of queries on the segmentation performance. The results are presented in Fig. 6. The number of queries is an important aspect of such models as each query represents a mask and ultimately defines the upper bound of recognizable masks in a 3D scene. In the case of Mask3D, the segmentation performance remains relatively stable with changes in the number of queries, whereas Segment3D experiences notable benefits from an increased number of queries. This is an interesting observation since the key difference between Segment3D and Mask3D is the training data, and it shows the benefit of automatically generated SAM masks which are more diverse than the manually annotated ground truth masks in ScanNet. Constrained by limited GPU memory, we were unable to assess performance with more than 400 queries. Nevertheless, the trend of the curves suggests the potential for further improved performance with increased query numbers.

Pre-Training with Additional Data. Since RGB-D data is available in abundance and masks can be automatically generated, it is natural to ask if pre-training on additional masks will further improve the overall performance. To that end, we perform a first experiment where we select additional frames from the training set of ScanNet++ dataset (to increase the variety of training data). Again, we sample frames at roughly 1 FPS resulting in 34k frames, in addition to the previous 76k frames of ScanNet. Table 4 shows an impressive performance boost of 14% AP. For future work, considering that this improvement is obtained simply by adding more automatically generated masks to the pre-training, our approach seems promising to train on even more data. It could even be plausible to train on internet images combined with monocular depth estimation, such as ZoeDepth [3], to compute the partial point clouds.

Pre-Training Dataset (# Frames)	**AP**	**AP**$_{50}$
ScanNet (76k)	12.0	22.7
ScanNet (76k), ScanNet++ (34k)	13.7 (+14%)	24.7 (+9%)

Table 4. **Performance Increase with Additional Training Data.**

Model	**Segmentor**	**ScanNet++**	**Replica**
OpenMask3D [44]	Mask3D [40]	15.0	18.0
OpenMask3D [44]	Segment3D (Ours)	17.7 (+18%)	18.7 (+4%)

Table 5. **Open-Set 3D Scene Understanding Scores.** Evaluation metric is average precision (AP) with IoU threshold of 50%.

4.3. Qualitative Results

Fig. 4 shows several representative examples of Segment3D segmentation results on the ScanNet++ dataset. As can be seen, the scenes are quite diverse, presenting multiple challenges such as clutter and a wide range of mask sizes. Despite these challenges, our model predicts quite accurate and well localized segmentation masks. For example, compared to Mask3D, our method is able to segment the finer-grained objects on top of the bed and the shelf. The masks of the computer screen and the chair in front of it are also less fragmented than predictions of the baselines.

4.4. Application: Open-Set Scene Understanding

A real-world application of our class-agnostic 3D segmentation method is open-vocabulary 3D scene understanding, as implemented by the recent OpenMask3D [44]. Given a 3D scene, a user can search for arbitrary objects via text-prompts (see Fig. 5). A core component of OpenMask3D is Mask3D which segments the scene into masks. Since Mask3D is trained on the closed-set of ScanNet, its masks are not truly class-agnostic or open-vocabulary. We therefore replace Mask3D with our class-agnostic Segment3D. We evaluate on the closed-set labels of Replica and ScanNet++ in Table 5 and report an improvement of up to 18%. Yet, this score serves only as an indicator, since most masks are not considered in this closed-set evaluation. See the appendix for a discussion and more details.

5. Conclusion and Discussion

We have presented Segment3D, a simple, yet powerful class-agnostic 3D segmentation model. The model is trained entirely on automatically generated masks from SAM, a foundation model for image segmentation. It is competitive and even outperforms existing 3D segmentation models that rely on hand-labeled 3D training scenes.

Indeed, this raises the question of whether hand-labeled 3D training datasets are as essential as they were once thought to be. Similarly, in the case of test labels, our qualitative comparison indicates that Segment3D sporadically identifies small object masks not annotated in the ground truth. Consequently, the full performance of Segment3D might not be accurately reflected in the scores.

Overall, the work shows the potential of foundation models as automatic training label generators, and our preliminary experiments seem to indicate that more data will further increase the segmentation performance for Segment3D.

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. 3

[2] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. ARKitScenes: A Diverse Real-World Dataset for 3D Indoor Scene Understanding using Mobile RGB-D Data. *arXiv preprint arXiv:2111.08897*, 2021. 3

[3] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-Shot Transfer by Combining Relative and Metric Depth. *arXiv preprint arXiv:2302.12288*, 2023. 8

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*, 2021. 3

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *ECCV*, 2020. 4

[6] 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. 2

[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. 3

[8] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical Aggregation for 3D Instance Segmentation. In *ICCV*, 2021. 1, 2

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-Attention Mask Transformer for Universal Image Segmentation. In *CVPR*, 2022. 4

[10] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In *NIPS*, 2021. 4

[11] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In *CVPR*, 2019. 4, 5

[12] 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. 3, 4, 5, 6

[13] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. BundleFusion: Real-time Globally Consistent 3D Reconstruction using On-the-fly Surface Re-integration. *TOG*, 2017. 2, 4

[14] 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

[15] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3D-MPA: Multi-Proposal Aggregation for 3D Semantic Instance Segmentation. In *CVPR*, 2020. 1

[16] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*, 1996. 5

[17] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient Graph-Based Image Segmentation. *IJCV*, 2004. 5, 6

[18] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-Vocabulary Object Detection via Vision and Language Knowledge Distillation. In *ICLR*, 2022. 2

[19] Huy Ha and Shuran Song. Semantic Abstraction: Open-World 3D Scene Understanding from 2D Vision-Language Models. In *CoRL*, 2022. 3

[20] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. CLIP2Point: Transfer CLIP to Point Cloud Classification with Image-Depth Pre-Training. In *ICCV*, 2023. 3

[21] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. KinectFusion: Real-time 3D Reconstruction and Interaction using a Moving Depth Camera. In *UIST*, 2011. 2, 4

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up Visual and Vision-Language Representation Learning with Noisy Text Supervision. In *ICML*, 2021. 2, 3

[23] 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. 2

[24] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language Embedded Radiance Fields. In *ICCV*, 2023. 2, 3

[25] 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. In *ICCV*, 2023. 1, 2, 3, 4

[26] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing Nerf for Editing via Feature Field Distillation. In *NIPS*, 2022. 3

[27] Zhihao Liang, Zhihao Li, Songcen Xu, Mingkui Tan, and Kui Jia. Instance Segmentation in 3D Scenes using Semantic Superpoint Tree Networks. In *CVPR*, 2021. 1, 2

[28] Jiahao Lu, Jiacheng Deng, Chuxin Wang, Jianfeng He, and Tianzhu Zhang. Query Refinement Transformer for 3D Instance Segmentation. In *ICCV*, 2023. 1, 2

[29] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-Vocabulary Point-Cloud Object Detection without 3D Annotation. In *CVPR*, 2023. 2, 3

[30] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *3DV*, 2016. 4

[31] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3D: Out-of-context data augmentation for 3D scenes. In *3DV*, 2021. 1

[32] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[33] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *ICCV*, 2021. 2

[34] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. OpenScene: 3D Scene Understanding with Open Vocabularies. In *CVPR*, 2023. 2, 3

[35] 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. 1, 2

[36] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, 2017. 1

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, 2021. 2, 3

[38] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-Guided Dense Prediction with Context-Aware Prompting. In *CVPR*, 2022. 2

[39] David Rozenberszki, Or Litany, and Angela Dai. Language-Grounded Indoor 3D Semantic Segmentation in the Wild. In *ECCV*, 2022. 5, 6

[40] 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, 2, 4, 5, 6, 7, 8

[41] Jonas Schult, Francis Engelmann, Theodora Kontogianni, and Bastian Leibe. DualConvMesh-Net: Joint Geodesic and Euclidean Convolutions on 3D Meshes. In *CVPR*, 2020. 1, 5

[42] 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. 3, 5

[43] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint Transformer for 3D Scene Instance Segmentation. In *AAAI*, 2023. 5

[44] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *NIPS*, 2023. 2, 3, 5, 7, 8

[45] 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

[46] 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, 2, 5

[47] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Towards Segmenting Everything in Context. In *ICCV*, 2023. 2

[48] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning Object Bounding Boxes for 3D Instance Segmentation on Point Clouds. In *NIPS*, 2019. 2

[49] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. SAM3D: Segment Anything in 3D Scenes. *arXiv preprint arXiv:2306.03908*, 2023. 5, 6, 7

[50] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. In *ICCV*, 2023. 2, 5

[51] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. GSPN: Generative Shape Proposal Network for 3D Instance Segmentation in Point Cloud. In *CVPR*, 2019. 2

[52] Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. CLIP2: Contrastive Language-Image-Point Pretraining from Real-World Point Cloud Data. In *CVPR*, 2023. 3

[53] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A Simple Framework for Open-Vocabulary Segmentation and Detection. In *CVPR*, 2023. 2

[54] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. PointCLIP: Point Cloud Understanding by CLIP. In *CVPR*, 2022. 3

[55] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized Decoding for Pixel, Image, and Language. In *CVPR*, 2023. 2

[56] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment Everything Everywhere All at Once. In *NIPS*, 2023. 2