CLIP-FO3D: Learning Free Open-world 3D Scene Representations from 2D Dense CLIP

Junbo Zhang[1] Runpei Dong[2] Kaisheng Ma[1]*
[1]Tsinghua University [2]Xi'an Jiaotong University
zhangjb21@mails.tsinghua.edu.cn, runpei.dong@stu.xjtu.edu.cn, kaisheng@mail.tsinghua.edu.cn

Abstract

Training a 3D scene understanding model requires complicated human annotations, which are laborious to collect and result in a model only encoding close-set object semantics. In contrast, vision-language pre-training models (e.g., CLIP) have shown remarkable open-world reasoning properties. To this end, we propose directly transferring CLIP's vision feature space to 3D scene understanding model without any form of supervision. We first modify CLIP's input and forwarding process so that it can be adapted to extract dense pixel features for 3D scene contents. We then project multi-view image features to the point cloud and train a 3D scene understanding model with feature distillation. Without any annotations or additional training, our model achieves promising annotation-free semantic segmentation results on open-vocabulary semantics and long-tailed concepts. Besides, serving as a cross-modal pre-training framework, our method can be used to improve data efficiency during fine-tuning. Our model outperforms previous SOTA methods in various zero-shot and data-efficient learning benchmarks. Most importantly, our model successfully inherits CLIP's rich-structured knowledge, allowing 3D scene understanding models to recognize not only object concepts but also open-world semantics.

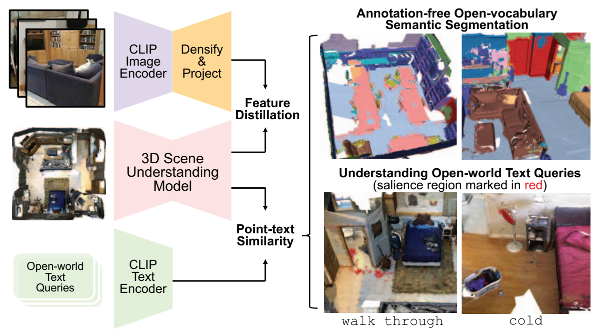

Figure 1. Overview of CLIP-FO3D. We directly transfer CLIP's vision feature space to 3D point representations by extracting dense pixel features, 2D-to-3D feature projection, and feature distillation. Without any human annotations, CLIP-FO3D successfully utilizes the point-text feature similarity to perform semantic segmentation on open-vocabulary semantics and understand open-world knowledge (*e.g.* affordances, colors, activities).

1. Introduction

3D scene understanding aims to distinguish objects' semantics, identify their locations, and infer the geometric attributes from 3D scene data. It has a wide range of applications in virtual reality [53], robot navigation [3, 51] and autonomous driving [38, 75, 76]. However, training traditional 3D scene understanding models requires a large number of human annotations, which are laborious to collect. Besides, the human annotations used in the current 3D scene understanding benchmark only contain close-set semantic information of the objects (*e.g.*, 20 classes in ScanNet [15]). These make it difficult for 3D scene understanding systems to recognize open-vocabulary categories and to infer open-world semantics.

Large-scale vision-language foundation models (*e.g.*, CLIP [57]) capture rich visual and language features. They only require image-text pairs mined from the Internet for unsupervised pre-training, and have demonstrated superior ability in zero-shot and open-vocabulary reasoning for classification and dense prediction tasks [25, 39, 79, 91]. However, due to the difficulty of collecting 3D-text pairs and the complexity of scene data, is extremely challenging to build analogous 3D foundation models.

Some pioneering works explore the potential of extending CLIP's text feature space to 3D scene representations. LSeg [39]/OpenSeg [25] models (used in OpenScene [54]) and CLIP2Scene [6] align the pixel/point features with class names' CLIP text embeddings. However, the text embeddings contain much less knowledge than CLIP's vision encoder, because some attributes, rare and abstract concepts would be difficult to label in the training texts. Besides, OpenScene [54] requires 2D segmentation annotations for the pre-trained models. CLIP2Scene [6] requires close-set pre-defined classes for training. This will result in a loss of CLIP's powerful open-world properties because training with pre-defined classes restricts the feature space to a lim-

*Corresponding Author.

ited set of vocabularies.

In this paper, we propose directly transferring CLIP's vision feature space to 3D scene understanding models without any supervision (*e.g.*, 2D/3D annotations or vision-language grounding annotations). Once the transfer is complete, our model can complete the open-vocabulary 3D scene understanding task without additional training. This annotation-free paradigm is both effort-saving and able to preserve CLIP's open-world information to the maximum.

In this direction, MaskCLIP [91] in 2D vision has taken the first step towards extracting pixel-level features from CLIP vision encoder's final feature map. However, it has difficult adapting to 3D contents since 3D scenes contain much more objects and more complex structures. Besides, MaskCLIP has an inherent defect in precisely locating and segmenting objects' boundaries due to the limited resolution of CLIP's final feature map. To address these problems, we propose a new method to extract free pixel-level CLIP features without breaking its feature space. We first crop the input images at multi-scale to accommodate various object sizes. To preserve object-level semantics, we divide each cropped sample into semantically relevant regions and extract a local feature for each region. Specifically, we add several local classification tokens in ViT [21] encoder, which only aggregate information from local patches within a region. By simply forwarding CLIP and combining multiple local features, we extract a pixel-level feature map for each 3D scene's RGB view.

After extracting pixel-level CLIP features, we adopt the feature projection in 3DMV [16] to project multi-view image features to the point cloud. The resulting point features are aligned with CLIP's feature space and used as off-line training targets. Then we train the 3D understanding model with feature distillation, minimizing the distance between learned point features and the target features as in [54]. This way, we obtain CLIP-FO3D, which extracts free and open-world 3D scene representations aligned with CLIP.

Our CLIP-FO3D can perform annotation-free open-vocabulary 3D semantic segmentation without additional training processes. Since CLIP's vision features are well aligned with text features, we can take the text embeddings of each class name's prompts as classification weights to perform semantic segmentation. CLIP-FO3D performs remarkably on standard close-set ScanNet [15] and S3DIS [4] segmentation benchmark. In addition, to examine CLIP-FO3D's open-vocabulary capability inherited from CLIP, we extend the label set of the standard ScanNet dataset with NYU labels [63]. We demonstrate that CLIP-FO3D has remarkable segmentation results on NYU-40 classes and other long-tailed categories beyond the NYU label set.

Besides, given that collecting 3D point clouds and annotating are laborious, CLIP-FO3D can also be regarded as an unsupervised cross-modal pre-training framework to benefit data efficiency. CLIP-FO3D achieves promising performance in traditional benchmarks where limited annotations are provided, such as zero-shot and data-efficient learning.

Most importantly, CLIP-FO3D encodes rich open-world knowledge inherited from CLIP. It understands not only object concepts, but also text queries with open-world semantics (*e.g.* affordances, colors, activities), broadening the application of 3D scene understanding.

Our contributions can be summarized as follows:

- We propose directly transferring CLIP's vision feature space to 3D scene understanding models without any annotation, which preserves CLIP's open-world properties to the maximum.
- We present a novel method to extract pixel-level CLIP features, and a feature distillation method to align 3D point representations with CLIP's feature space.
- Our model achieves promising annotation-free 3D semantic segmentation performance on large vocabularies, and shows remarkable open-world properties.
- As an unsupervised pre-training method, our model outperforms previous state-of-the-art methods in zero-shot and data-efficient learning.

2. Related Work

2.1. Open-vocabulary Dense Prediction in 2D Vision

Open-vocabulary dense prediction aims to recognize and localize objects with open-set semantics. Pioneering works [43, 71, 81, 92] in 2D vision mainly utilize large-scale image-caption annotations as weak supervision source to enlarge the vocabulary set. Other works [22, 23, 25, 26, 39, 41, 47, 49, 58, 70, 80, 90] distill vision-language foundation models' (*e.g.*, CLIP) knowledge to transfer their open-vocabulary capability. However, they all require some form of human annotations, such as box/mask proposals and pixel semantic annotations. Open-world knowledge encoded in CLIP is *forgotten* when fine-tuning with human annotations. In contrast, MaskCLIP [91] proposes directly utilizing CLIP for dense prediction tasks without training. However, it struggles to handle 3D scene's RGB views with more complex contents, as shown in Section 4. We propose a new method to extract free pixel-level CLIP features from 3D scene's RGB views by only modifying CLIP's inputs and forwarding process *without* fine-tuning.

2.2. Zero-shot 3D Visual Recognition

Zero-shot learning is relatively under-explored in 3D vision. Initial works focus on classification tasks on object-level data [10–13, 29, 55, 72]. Recent works [33, 42, 83, 86] adopt CLIP to perform zero-shot 3D classification with open-vocabularies. Multi-modal foundation models are also

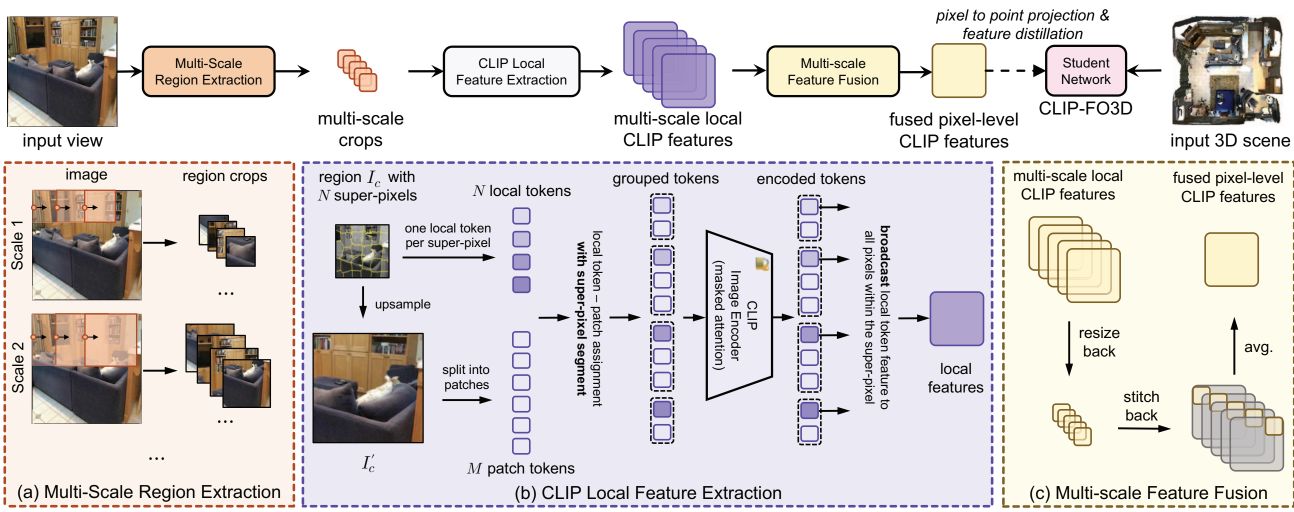

Figure 2. **Top:** Overall training process for CLIP-FO3D. For RGB views, we extract free pixel-level CLIP features and project pixel features to 3D points. For 3D scenes, we encode the point features and train CLIP-FO3D with feature distillation. **Bottom:** Three steps to extract pixel-level CLIP features. **(a)** The input image is cropped at multi-scale to accommodate various object sizes. **(b)** Extracting pixel-level feature map for each region crop. Each region crop I_c is divided into N super-pixels, and a local token is initialized for each super-pixel. The crop is then upsampled and input into CLIP's ViT encoder together with the local tokens. Each patch token in ViT is assigned to a local token based on the spatial location of patch and super-pixel. The attention score of local token is computed only from its assigned patches. The encoded local token feature is finally broadcast to all pixels within the super-pixel. **(c)** All region crops' feature maps are resized back and stitched to the original image. We average the multiple features of pixels to obtain final pixel-level CLIP features.

used in other 3D tasks, such as NeRF [35, 37, 65] and 3D visual question answering [17].

For 3D scene understanding, Michele *et al*. [48] and Chen *et al*. [7] study zero-shot semantic segmentation with generative model and word embedding prototypes. Lu *et al*. [46] propose a zero-shot 3D object detection method with pseudo-labels generated with 2D classifier. Zhang *et al*. [89], Chen *et al*. [6] and Rozenberszki *et al*. [61] propose fully unsupervised 3D semantic/instance segmentation methods. But these methods still result in a model with close-set vocabularies. Recent works achieve promising results in applying CLIP to open-vocabulary 3D scene understanding. However, they all require human annotations for training, such as 2D pixel labels [54], 3D point labels [27] and pre-trained masks [34, 64]. PLA [18] and RegionPLC [73] utilize an image-captioning model to generate captions for 3D content, enabling open-vocabulary segmentation by aligning 3D representations with text embeddings. Unlike existing works, we propose directly transferring CLIP's vision feature space to 3D scene representations with no annotations. Without additional training, CLIP-FO3D shows superior open-world understanding properties inherited from CLIP.

2.3. 3D Representation Learning

Inspired by the self-supervised representation learning in 2D vision, 3D pre-training methods [30, 82] achieve better fine-tuning performance and efficiency in various downstream tasks by leveraging contrastive learning [8, 9, 44, 68, 88], masked auto-encoder [20, 52, 78, 85], or both [55].

Beyond self-supervised pre-training, cross-modal learning methods propose to distill knowledge in images/text and pre-trained models to 3D representations [2, 20, 32, 55, 59]. For scene understanding tasks, recent works enrich 3D point representations by utilizing 2D/text annotations [50, 77, 84], pixel-point alignments [40, 45, 62, 66], neural rendering [31] and pre-trained models [56, 60, 69, 74, 87]. Our method can be regarded as an unsupervised cross-modal 3D representation learning method. We demonstrate that distilling CLIP's richly-structured vision knowledge to 3D models can benefit 3D scene understanding when limited annotations are available, outperforming previous self-supervised and cross-modal pre-training methods.

3. Method

This section describes the process of transferring CLIP's feature space to 3D scene understanding models. We first extract pixel-level CLIP features from the 3D scene's RGB views by modifying CLIP's inputs and forwarding process, introduced in Section 3.1. We then get target point features from pixel features and train the 3D model by feature distillation, introduced in Section 3.2.

3.1. Extract Free Pixel-level CLIP Features

Although CLIP only aligns image-level global features with text embeddings, it should inherently encode local and dense semantics at the front layers. As demonstrated in MaskCLIP [91], CLIP must divide image-level semantics into local segments, and properly align each segment's semantics with independent concepts in the text. MaskCLIP proposes to discard the global pooling layer and extract dense features from the final feature map with reformulated 1×1 convolutional layers.

However, adapting MaskCLIP to 3D scene contents brings poor dense prediction results, as shown in Section 4. On the one hand, CLIP's feature map has a much lower resolution than the input images (downsampled by 16^2 in ViT/16 [21] and 32^2 in ResNet-50 [28]). Although MaskCLIP is reasonably capable of recognizing salient semantics in the images, it struggles to segment the numerous objects in 3D scenes at different scales. On the other hand, the pixel features in CLIP's final feature map contain much global semantics regarding the entire image. They may contain multiple objects' semantics since each pixel's feature aggregates information from all other pixels in forwarding. However, ideally, we hope to extract object-level features of different objects in a 3D scene. We propose increasing the resolutions and extracting local features from CLIP to address the aforementioned problems, described as follows.

Multi-scale region extraction. Firstly, the input view is cropped at multi-scales to adapt to the recognition of objects of various sizes in 3D scenes, as shown in Figure 2 (a). For each cropped sample, we divide into many local regions. This effectively improves the feature resolution. Specifically, we divide the image sample into super-pixels with SLIC [1] as in [62]. The super-pixel roughly covers an object or object's part, resulting in locally visually similar regions. After processing the input image, we extract an embedding vector for each super-pixel with modified CLIP's ViT [21] encoder, which is introduced below.

Extracting local features from CLIP. We hope the super-pixel feature aggregates information from local patches rather than from the entire image like the global classification token in ViT's encoder. This process is shown in Figure 2 (b). Given a cropped image sample I_{c}, we segment it into N super-pixels: $I_{\mathrm{c}} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots \cup \mathcal{S}_N$, where $\mathcal{S}_i \cap \mathcal{S}_j = \varnothing,\ \forall\, i \neq j$. We then upsample the exact cropped image to CLIP's input size and obtain $I_{\mathrm{c}}^{'}$. In ViT, $I_{\mathrm{c}}^{'}$ is first reshaped into M flattened patches: $I_{\mathrm{c}}^{'} = \mathcal{P}_1 \cup \mathcal{P}_2 \cup \cdots \cup \mathcal{P}_M$, where $M = 14^2$ in ViT-B/16. We then assign each patch in $I_{\mathrm{c}}^{'}$ to a specific super-pixel in I_{c} based on the their spatial locations by interpolation, since the number of super-pixels N in our method is always less than the number of patches M. We denote the patch $\mathcal{P}_i$ being assigned to the super-pixel $\mathcal{S}_j$ as $\mathcal{P}_i \sim \mathcal{S}_j$.

To represent each super-pixel's feature, we add N local classification tokens beyond the original global classification token in CLIP's forwarding process. The local tokens have similar functions with the group tokens in [71], but are not learnable and have different updating mechanism. The local tokens are initialized the same as the global one and are updated by the same self-attention mechanism and pre-trained weights in ViT. The only difference in updating the local tokens during inference is how attention scores are computed. Namely, each local token is only attended to its assigned patch tokens, which is implemented as masked-attention mechanism as in [19, 71]. Recall that in ViT's forwarding process, the attention score of the classification token is calculated as:

$$\begin{aligned} A_{\text{global}} &= \sum_i \operatorname{softmax}\left(\frac{q^{\mathrm{g}} \cdot k_i^{\mathrm{T}}}{C}\right) v_i, \\ q^{\mathrm{g}} =& \operatorname{Emb}_{\mathrm{q}}(x^{\mathrm{g}}),\ \ k_i = \operatorname{Emb}_{\mathrm{k}}(x_i),\ \ v_i = \operatorname{Emb}_{\mathrm{v}}(x_i), \end{aligned} \tag{1}$$

where C is a constant scaling factor and $\operatorname{Emb}(\cdot)$ denotes the linear layers encoding the query, key, and value embeddings. x^{g} is the *global* classification token and x_i represents the input feature of patch $\mathcal{P}_i$.

In our method, the attention score of each local classification token is computed from local patches as:

$$\begin{aligned} A_{\text{local},\, j} &= \sum_{i:\, \mathcal{P}_i \sim \mathcal{S}_j} \operatorname{softmax}\left(\frac{q_j^{\ell} \cdot k_i^{\mathrm{T}}}{C}\right) v_i, \\ q_j^{\ell} =& \operatorname{Emb}_{\mathrm{q}}(x_j^{\ell}),\ k_i = \operatorname{Emb}_{\mathrm{k}}(x_i),\ v_i = \operatorname{Emb}_{\mathrm{v}}(x_i), \end{aligned} \tag{2}$$

where $\mathcal{P}_i \sim \mathcal{S}_j$ means that patch $\mathcal{P}_i$ is assigned to the super-pixel $\mathcal{S}_j$. x_j^{ℓ} is the *local* classification token of super-pixel $\mathcal{S}_j$ and x_i represents the input feature of patch $\mathcal{P}_i$.

Note that the original tokens in ViT are not affected during inference. By forwarding CLIP with additional tokens, we obtain a local feature for each super-pixel that is aligned with CLIP's feature space. We then apply the same local token feature to all pixels within a super-pixel to preserve object-level semantics. In this way, we obtain a feature map for each cropped image of the same size as CLIP's input.

Multi-scale feature fusion. After extracting local feature maps for all the cropped samples, we resize them back to the cropping sizes and stitch them back to the original input image, as shown in Figure 2 (c). Since one pixel may belong to different super-pixels from different cropped samples, we average all features as the final pixel feature. We extract pixel features for each 3D scene's RGB view. While the whole process is slow and difficult to apply to real-time inference, we only do the above process once for each view and use the pixel-level features as offline training targets.

3.2. Feature Distillation with 2D Teacher

To connect the pixel-level 2D features with 3D point features, we project each point in a 3D scene back to the RGB views following 3DMV [16]. Point-to-pixel projection is computed based on the camera pose, intrinsics, and the world-to-grid transformation matrix. Since we can obtain each projected pixel's depth value from the RGB-D images, we only keep the points that are in the correct depth ranges

	Backbone	ScanNet		S3DIS	
		mIoU	mAcc	mIoU	mAcc
Inference by feature projection					
MaskCLIP-3D	CLIP	9.7	21.6	-	-
Target Feature (Ours)	CLIP	27.6	47.7	-	-
CLIP-FO3D (Ours)	Res-Unet	**30.2**	**49.1**	**22.3**	**32.8**

Table 1. Annotation-free semantic segmentation on standard ScanNet and S3DIS datasets. *MaskCLIP-3D* and our *Target Feature* use pixel features to infer points' semantics, which is slow and unsuitable for practical use. Our pixel-level target features capture significantly better semantics than MaskCLIP. CLIP-FO3D trained with feature distillation even outperforms its target.

	Head		Common		Tail	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
Inference by feature projection						
MaskCLIP-3D	19.8	33.1	13.3	26.8	7.5	15.3
Target Feature (Ours)	37.1	63.4	**39.0**	**55.4**	**40.6**	**55.3**
CLIP-FO3D (Ours)	**44.3**	**65.9**	37.6	50.6	26.5	36.4

Table 2. Open-vocabulary semantic segmentation on ScanNet with extended labels from NYU label set. We divide all categories into *Head*, *Common* and *Tail* based on the point numbers.

for further training. As some points will be associated with multiple pixels from different views, we compute the average of all 2D features as the final point feature $f_{3D} \in \mathbb{R}^C$, where the channel number C is the same as CLIP's vision encoder. As a result, we obtain point-level features for each 3D scene denoted as $\mathcal{F}_{\text{target}} = \{\hat{f}_{3D,i}\}_{i=1}^{N_p}$, where N_p is the number of points in the scene.

After extracting offline target point features for each scene, we train the 3D understanding model to learn from these targets by feature distillation like OpenScene [54]. Denoting the learned point features for a scene as $\mathcal{F}_{\text{learn}} = \{f_{3D,i}\}_{i=1}^{N_p}$, the loss function of feature distillation is:

$$\mathcal{L} = \frac{1}{N_p} \sum_{i}^{N_p} \mathcal{D}\left(f_{3D,i},\ \hat{f}_{3D,i}\right). \tag{3}$$

The distance $\mathcal{D}(\cdot,\cdot)$ is the negative cosine similarity:

$$\mathcal{D}(f_1, f_2) = -\frac{f_1}{\|f_1\|_2} \cdot \frac{f_2}{\|f_2\|_2}, \tag{4}$$

where $\|\cdot\|_2$ is $\mathcal{L}_2$-norm. We use cosine distance because CLIP-driven classification relies on the cosine distance between vision and text embeddings.

The whole training process only requires the 3D dataset and the pre-trained CLIP vision encoder, without any form of supervision (*e.g.*, 2D/3D annotations or vision-language grounding annotations). Since the learned point features are consistent with CLIP's feature space, the 3D model can perform open-vocabulary semantic segmentation and open-world reasoning once the feature distillation is finished.

4. Experiments

4.1. Experimental Setup

Dataset. We train CLIP-FO3D on ScanNet's training set [15]. We use the RGB-D images and the 3D scene meshes for training, and no labels are used. We sub-sample RGB-D images from the raw ScanNet videos every ten frames for each scene. We evaluate our method on ScanNet [15] and S3DIS [4] datasets. ScanNet's validation set and S3DIS's "Area 5" are used for all the evaluation experiments. Notice that raw categories' names and the mapping to NYU40 label set are provided in the ScanNet dataset, which we use to enlarge the vocabulary. We remove the "other furniture" category in ScanNet and the "clutter" category in S3DIS because they do not have specific semantics that can be classified with any text embeddings.

Implementation Details. We use CLIP's ViT-B/16 encoder as the 2D backbone and use the corresponding text encoder for generating text embeddings. For all the experiments, we adopt MinkowskiNet16 [14] as 3D scene understanding backbone. We remove the final classifier and change the output feature dimension to 512 to match CLIP's feature dimension. We train CLIP-FO3D using SGD optimizer with a learning rate of 0.8 and a batch size of 4. The model is trained for 80K steps, and the learning rate is decreased by 0.99 for every 1,000 steps. The fine-tuning experiments on CLIP-FO3D are trained with a batch size of 4 for 40K steps. We set the initial learning rate of the pre-trained CLIP-FO3D networks to 0.001, with polynomial decay with power 0.9, because we find that a small learning rate leads to better results when fine-tuning CLIP's feature space. The initial learning rate of the classifier in data-efficient learning is set to 0.1 following the original setting in [30]. For zero-shot learning and data-efficient learning experiments, we follow the original benchmarks in [7] and [30]. More implementation details can be found in the supplementary material.

4.2. Annotation-free Semantic Segmentation

Semantic segmentation on standard benchmarks. We first present the results of annotation-free semantic segmentation on standard ScanNet and S3DIS benchmarks, where no supervision is provided during training. This is a rather challenging setting. After training CLIP-FO3D with feature distillation, we can take the text embeddings of each class name's prompts as classification weights to perform semantic segmentation.

The results are shown in Table 1. The first two rows show the results of using multi-view pixel features to infer each point's semantics by feature projection (introduced in Section 3.2). ***MaskCLIP-3D*** is a baseline method that we use MaskCLIP to compute pixel features for RGB views as in [91]. ***Target Feature*** is our proposed method to extract free dense CLIP features from RGB views introduced

Setting	Unseen-2			Unseen-4			Unseen-6			Unseen-8			Unseen-10		
Metric	mIoU $\mathcal{S}$	mIoU $\mathcal{U}$	hIoU	mIoU $\mathcal{S}$	mIoU $\mathcal{U}$	hIoU	mIoU $\mathcal{S}$	mIoU $\mathcal{U}$	hIoU	mIoU $\mathcal{S}$	mIoU $\mathcal{U}$	hIoU	mIoU $\mathcal{S}$	mIoU $\mathcal{U}$	hIoU
3DGenZ [48]	33.4	12.8	18.5	32.8	7.7	12.5	31.2	4.8	8.3	29.7	2.1	3.9	30.1	1.4	2.7
TGP [7]	58.6	51.6	54.9	57.9	34.1	42.9	55.2	15.4	24.1	51.6	12.1	19.6	52.5	9.5	16.1
MaskCLIP-3D	67.3	38.7	49.1	64.4	31.0	41.9	62.9	24.7	35.5	61.8	19.1	29.2	62.3	13.4	22.1
CLIP-FO3D (Ours)	**70.6**	**60.3**	**65.0**	**69.5**	**54.8**	**61.2**	**67.3**	**50.8**	**57.9**	**65.8**	**46.7**	**54.6**	**67.7**	**40.7**	**50.8**

Table 3. Zero-shot semantic segmentation on ScanNet with different settings. "Unseen-i" indicates that there are i classes that have no labels during training. $\mathcal{S}$ and $\mathcal{U}$ represent the performance of seen and unseen classes, respectively.

Setting	Unseen-4			Unseen-6		
Metric	mIoU $\mathcal{S}$	mIoU $\mathcal{U}$	hIoU	mIoU $\mathcal{S}$	mIoU $\mathcal{U}$	hIoU
3DGenZ [48]	53.1	7.3	12.9	28.3	6.9	11.1
TGP [7]	60.4	20.6	30.7	-	-	-
CLIP-FO3D (Ours)	**64.8**	**26.1**	**37.2**	**60.8**	**29.3**	**39.5**

Table 4. Zero-shot semantic segmentation on S3DIS with different settings. "Unseen-i" indicates that there are i classes that have no labels during training. $\mathcal{S}$ and $\mathcal{U}$ represent the performance of seen and unseen classes, respectively.

in Section 3.1. Notice that one should forward the vision encoder for all RGB views and then align 3D points with pixels with camera pose, intrinsics and transformation matrix. This inference process by feature projection is very slow and not applicable for practical use. In contrast, CLIP-FO3D is convenient for processing new 3D data. However, the results can still be compared and give some inspiration.

It is observed that our Target Features extract more reasonable semantics than MaskCLIP-3D, indicating that the proposed method is indispensable in extracting dense CLIP features for 3D contents. Moreover, our CLIP-FO3D achieves remarkable segmentation results on ScanNet and S3DIS in this challenging setting (notice that four categories in S3DIS never appear during the training on ScanNet). CLIP-FO3D even outperforms its arget features in ScanNet, indicating that some misleading target features can be corrected during feature distillation.

Semantic segmentation with open vocabularies. The standard ScanNet benchmark only contains a small vocabulary of 20 classes. To examine the open-vocabulary capability of CLIP-FO3D inherited from CLIP, we first extend the original vocabulary size with the NYU-40 label set. We remove the NYU-40 labels that do not have specific semantics (*e.g.*, "other structure", "other furniture", "other prop") and evenly divide all the rest categories into *Head*, *Common* and *Tail* based on the point numbers of each category. We show the semantic segmentation results on the three category sets in Table 2. Our Target Features outperform the baseline MaskCLIP-3D by a large margin. Since categories in *Common* and *Tail* set usually contain objects with smaller sizes (*e.g.*, bag, box, pillow, book), MaskCLIP-3D struggles to extract their semantics due to the limited feature resolution. While our method even extracts higher quality features for *Tail* categories than *Head*, thanks to the multi-scale inputs and local super-pixel features. CLIP-FO3D also achieves promising results in all categories, while it does not perform as well as our Target Features for *Common* and *Tail* categories, due to the limited number of examples in the training set. However, CLIP-FO3D can be easily used for indoor open-vocabulary scene understanding applications.

Qualitative Results. We show the visualization of annotation-free semantic segmentation results on standard ScanNet benchmark and recognizing long-tailed categories. As shown in Figure 3 (left), decent segmentation masks are obtained compared to the ground truth. Our method differs from ground-truth results for some objects with ambiguous semantics, such as `table` and `desk`, `cabinet` and its `door`, `chair` and `sofa`. In Figure 3 (right), our method successfully recognizes long-tailed categories that are not annotated in traditional benchmarks, demonstrating our model's promising open vocabulary capability.

4.3. Zero-shot Semantic Segmentation

Zero-shot semantic segmentation methods train the 3D scene understanding model only with the labels on a subset of classes (seen classes), and evaluate both seen classes and unseen classes. All the existing methods in 3D scene understanding use transductive settings, in which the unlabeled points are accessible during training. CLIP-FO3D can be applied to zero-shot semantic segmentation with minor effort. Specifically, CLIP-FO3D can be used to generate pseudo-labels for the unlabeled points.

We compare our method with state-of-the-art zero-shot semantic segmentation methods on ScanNet and S3DIS following the benchmarks in [7]. We also use a new benchmark where six classes are chosen as unseen classes in S3DIS. We use the metric of mean intersection over union (mIoU) for both seen classes ($\mathcal{S}$) and unseen classes ($\mathcal{U}$), and use the harmonic mean IoU (hIoU) to demonstrate the overall performance of zero-shot learning as in [5, 67].

As shown in Table 3, our method outperforms the previous state-of-the-art method [7] by 10.1%, 18.3%, 33.8%, 35.0% and 34.7% on hIoU metric when there are 2, 4, 6, 8 and 10 unseen classes during training. Our CLIP-FO3D also outperforms the method of pseudo-labeling with

Figure 3. Qualitative results of annotation-free semantic segmentation on ScanNet and recognizing long-tailed categories.

Method	Sup.	Percentage of labeled scene 1%	5%	10%	20%
Scratch	-	28.1	47.4	58.9	63.8
CSC [30]	✗	29.7	48.8	59.9	64.7
LangGround [60]	✓	31.9	49.3	60.8	65.3
CLIP-FO3D (Ours)	✗	**36.1** (↑8.0)	**51.7** (↑4.3)	**63.1** (↑4.2)	**66.8** (↑3.0)

Table 5. Data-efficient semantic segmentation on ScanNet with **limited scene reconstructions**. "Sup." indicates whether points' semantic labels are used during pre-training.

Method	Sup.	Number of labeled points 20	50	100	200
Scratch	-	46.3	58.3	62.8	65.4
CSC [30]	✗	54.0	60.7	65.6	68.3
LangGround [60]	✓	55.1	62.4	66.0	68.2
CLIP-FO3D (Ours)	✗	**57.6** (↑11.3)	**64.3** (↑6.0)	**68.2** (↑5.4)	**69.5** (↑4.1)

Table 6. Data-efficient semantic segmentation on ScanNet with **limited point annotations**. "Sup." indicates whether points' semantic labels are used during pre-training.

MaskCLIP-3D by large margins. It is observed that our method is more effective than the baselines when there are more unseen classes, demonstrating superior zero-shot capability. The results on S3DIS with two settings in Table 4 reflect a similar phenomenon, although S3DIS's data is not accessible during the training of CLIP-FO3D.

4.4. Data-efficient 3D Scene Understanding

As the collection and annotation of 3D point cloud data are very laborious, data-efficient learning methods have been proposed to train a better 3D model when training data or labels are scarce. Existing works have explored self-supervised [30] or cross-modal [60] pre-training methods for data-efficient fine-tuning. CLIP-FO3D can also be considered an unsupervised cross-modal pre-training method to benefit data efficiency. Unlike existing works, our method leverages the richly-structured feature space inherited from CLIP to improve data-efficient fine-tuning results.

Method	Annotation-free mIoU	mAcc	Zero-shot (Unseen-4) $\mathcal{S}$	$\mathcal{U}$	hIoU
w/o Multi-scale	17.1	33.8	68.1	43.3	52.9
w/o Local Feature	21.6	39.3	68.3	48.5	56.7
CLIP-FO3D (Ours)	**30.2**	**49.1**	**69.5**	**54.8**	**61.2**

Table 7. Ablation study on ScanNet for two benchmarks. "Multi-scale" represents cropping input images at multi-scales when extracting pixel-level features with CLIP. "Local Feature" represents using local classification tokens to extract super-pixel features.

We follow the official data-efficient learning benchmarks [30]: Limited Reconstructions (only a few labeled scenes are used for training) and Limited Annotations (only a few points are labeled in each scene). "Scratch" denotes the training from scratch baseline. We also remove the "other furniture" category in ScanNet as in the annotation-free segmentation experiments. The results of limited scene reconstructions are shown in Table 5, when only 1%, 5%, 10%, and 20% scenes are used during training. Our method achieves mIoU improvements of 8.0%, 4.3%, 4.2% and 3.0% over training from scratch, and achieves superior results compared to the previous state-of-the-art supervised pre-training method. The results of limited point annotations are shown in Table 6, when 20, 50, 100, and 200 annotated points are randomly sampled for each scene. Our method achieves mIoU improvements of 11.3%, 6.0%, 5.4%, and 4.1% over training from scratch, and outperforms previous state-of-the-art supervised pre-training methods.

5. Discussion

Ablation study. Table 7 shows the ablation study on the proposed methods when extracting pixel features from CLIP. "Multi-scale" represents increasing the feature resolution with multi-scale inputs, and "Local Feature" represents extracting super-pixel features using local classification tokens. We show the results of annotation-free and zero-shot semantic segmentation (with four unseen classes) on ScanNet. It is observed that both techniques are crucial for transferring CLIP's feature space. While each single

Figure 4. Open-world 3D scene understanding on ScanNet. Texts are used as queries and the most relevant areas are marked in red.

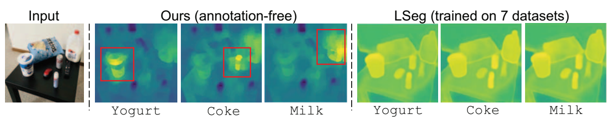

Figure 5. Pixel-text similarities of our method and LSeg [39]. Colors from blue to yellow indicate the similarity from low to high.

method outperforms the MaskCLIP-3D baseline, increasing the feature resolution with multi-scale inputs brings more significant improvements than extracting local features.

Open-world 3D scene understanding. Since CLIP is trained with massive image-text pairs mined from the Internet, it inherently encodes rich real-world knowledge which can guide open-world applications such as visual navigation and embodied AI [24, 36]. Since we directly transfer CLIP's feature space to 3D models, CLIP-FO3D has the potential to accomplish open-world 3D scene understanding.

Inspired by the qualitative results in [47] and [54], we use text embeddings to query open-world semantics for CLIP-FO3D's scene representations. Given a text description, we extract its feature with CLIP's text encoder, calculate its similarity with point features, and then threshold to produce a 3D mask. The visualization results are shown in Figure 4. Our model successfully finds the location that is most relevant to the text descriptions. For example, given "`write`", CLIP-FO3D finds the whiteboard and desk on which one can write. Our model is also able to understand other color, activity and affordance descriptions. These results demonstrate that our 3D scene representations encode rich and well-structured knowledge about the real world.

Advantages of Annotation-free Training. We demonstrate that annotation-free training preserves CLIP's rich information to the maximum. In contrast, fine-tuning CLIP with human annotations restricts the feature space, as demonstrated in [34].

We compare our pixel-level feature extraction with a supervised 2D segmentation method, LSeg [39]. LSeg aligns pixel embeddings to the CLIP text embedding of the corresponding semantic class from 7 segmentation datasets. Visualization results of capturing long-tailed concepts are shown in Figure 5. The example picture is used in [34]. The similarity of the pixels and text queries are used for visualization. Our method has outstanding advantages of recognizing long-tailed concepts directly inherited from CLIP. LSeg produces non-discriminative results because its feature space is restricted to pre-trained concepts and fails to generalize to open-world understanding.

6. Conclusion

This paper proposes transferring CLIP's feature space to 3D scene representations. We first extract pixel-level features from CLIP for RGB views. Then we adopt feature projection to get the target point features and train a 3D model. CLIP-FO3D achieves promising annotation-free semantic segmentation results on open-vocabulary concepts. It also outperforms previous state-of-the-art methods in zero-shot and data-efficient learning tasks. Moreover, our model inherits CLIP's open-world properties, encoding open-world knowledge beyond object concepts.

Acknowledgement

This research was partially supported by National Key R&D Program of China (2022YFB2804103), Key Research and Development Program of Shaanxi (2021ZDLGY01-05), Tsinghua University Dushi Program, National Natural Science Foundation of China (20211710187), and Tsinghua University Talent Program.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurélien Lucchi, Pascal V. Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 34:2274–2282, 2012. 4

[2] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 3

[3] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Jayant Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego M Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, and Mengyuan Yan. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv*, abs/2204.01691, 2022. 1

[4] Iro Armeni, Ozan Sener, Amir Roshan Zamir, Helen Jiang, Ioannis K. Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016. 2, 5

[5] Max Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2019. 6

[6] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Y. Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 1, 3

[7] Runnan Chen, Xinge Zhu, Nenglun Chen, Wei Li, Yuexin Ma, Ruigang Yang, and Wenping Wang. Zero-shot point cloud segmentation by transferring geometric primitives. *arXiv*, abs/2210.09923, 2022. 3, 5, 6

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

[9] Yujin Chen, Matthias Nießner, and Angela Dai. 4dcontrast: Contrastive learning with dynamic correspondences for 3d scene understanding. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 3

[10] Ali Cheraghian, Shafin Rahman, Dylan Campbell, and Lars Petersson. Mitigating the hubness problem for zero-shot learning of 3d objects. In *Brit. Mach. Vis. Conf. (BMVC)*, 2019. 2

[11] Ali Cheraghian, Shafin Rahman, Dylan Campbell, and Lars Petersson. Transductive zero-shot learning for 3d point cloud classification. In *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2019.

[12] Ali Cheraghian, Shafin Rahman, Townim Faisal Chowdhury, Dylan Campbell, and Lars Petersson. Zero-shot learning on 3d point cloud objects and beyond. *Int. J. Comput. Vis. (IJCV)*, 130:2364 – 2384, 2021.

[13] Ali Cheraghian, Shafin Rahman, and Lars Petersson. Zero-shot learning of 3d point cloud objects. In *International Conference on Machine Vision Applications (MVA)*, 2019. 2

[14] Christopher Bongsoo Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019. 5

[15] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017. 1, 2, 5

[16] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Eur. Conf. Comput. Vis. (ECCV)*, 2018. 2, 4

[17] Alexandros Delitzas, Maria Parelli, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. Multi-clip: Contrastive vision-language pre-training for question answering tasks in 3d scenes. *ArXiv*, abs/2306.02329, 2023. 3

[18] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 3

[19] Zheng Ding, Jie Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023. 4

[20] Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jian-Yuan Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 3

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent. (ICLR)*, 2021. 2, 4

[22] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guo Chun Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 2

[23] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 2

[24] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Clip on wheels: Zero-shot object navigation as object localization and exploration. *arXiv*, abs/2203.10421, 2022. 8

[25] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 1, 2

[26] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *Int. Conf. Learn. Represent.*

(ICLR), 2021. 2

[27] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models. In *Annu. Conf. Robot. Learn. (CoRL)*, 2022. 3

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016. 4

[29] Georg Hess, Adam Tonderski, Christoffer Petersson, Lennart Svensson, and Kalle AAstrom. Lidarclip or: How i learned to talk to point clouds. *arXiv*, abs/2212.06858, 2022. 2

[30] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021. 3, 5, 7

[31] Di Huang, Sida Peng, Tong He, Xiaowei Zhou, and Wanli Ouyang. Ponder: Point cloud pre-training via neural rendering. *arXiv*, abs/2301.00157, 2022. 3

[32] Rui Huang, Xuran Pan, Henry Zheng, Haojun Jiang, Zhifeng Xie, Shiji Song, and Gao Huang. Joint representation learning for text and 3d point cloud. *arXiv*, abs/2301.07584, 2023. 3

[33] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson W. H. Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. *arXiv*, abs/2210.01055, 2022. 2

[34] Krishna Murthy Jatavallabhula, Ali Kuwajerwala, Qiao Gu, Mohd. Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Varma Keetha, Ayush Kumar Tewari, Joshua B. Tenenbaum, Celso M. de Melo, M. Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. 2023. 3, 8

[35] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. *ArXiv*, abs/2303.09553, 2023. 3

[36] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021. 8

[37] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 3

[38] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019. 1

[39] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *Int. Conf. Learn. Represent. (ICLR)*, 2022. 1, 2, 8

[40] Zhenyu Li, Zehui Chen, Ang Li, Liangji Fang, Qinhong Jiang, Xianming Liu, Junjun Jiang, Bolei Zhou, and Hang Zhao. Simipu: Simple 2d image and 3d point cloud unsupervised pre-training for spatial-aware visual representations. In *AAAI Conf. Artif. Intell. (AAAI)*, 2022. 3

[41] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Péter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. *arXiv*, abs/2210.04150, 2022. 2

[42] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, H. Cai, Fatih Murat Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *ArXiv*, abs/2305.10764, 2023. 2

[43] Quande Liu, Youpeng Wen, Jianhua Han, Chunjing Xu, Hang Xu, and Xiaodan Liang. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 2

[44] Yunze Liu, Qingnan Fan, Shanghang Zhang, Hao Dong, Thomas A. Funkhouser, and Li Yi. Contrastive multimodal fusion with tupleinfonce. In *Int. Conf. Comput. Vis. (ICCV)*, 2021. 3

[45] Yueh-Cheng Liu, Yu-Kai Huang, HungYueh Chiang, Hung-Ting Su, Zhe-Yu Liu, Chin-Tang Chen, Ching-Yu Tseng, and Winston H. Hsu. Learning from 2d: Pixel-to-point knowledge transfer for 3d pretraining. *arXiv*, abs/2104.04687, 2021. 3

[46] Yuheng Lu, Chenfeng Xu, Xi Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary 3d detection via image-level class and debiased cross-modal contrastive learning. *arXiv*, abs/2207.01987, 2022. 3

[47] Timo Lüddecke and Alexander S. Ecker. Image segmentation using text and image prompts. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021. 2, 8

[48] Bjorn Michele, Alexandre Boulch, Gilles Puy, and Renaud Marlet. Generative zero-shot learning for semantic segmentation of 3d point clouds. In *Int. Conf. 3D Vis. (3DV)*, 2021. 3, 6

[49] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip H. S. Torr, and Ser Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. *arXiv*, abs/2212.04994, 2022. 2

[50] Yinyu Nie, Angela Dai, Xiaoguang Han, and Matthias Nießner. Learning 3d scene priors with 2d supervision. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 3

[51] Yong-Joo Oh and Y. Watanabe. Development of small robot for home floor cleaning. In *Proceedings of the 41st SICE Annual Conference. SICE*, 2002. 1

[52] Yatian Pang, Wenxiao Wang, Francis E. H. Tay, W. Liu, Yonghong Tian, and Liuliang Yuan. Masked autoencoders for point cloud self-supervised learning. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 3

[53] Youngmin Park, Vincent Lepetit, and Woontack Woo. Multiple 3d object tracking for augmented reality. In *IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 117–120, 2008. 1

[54] Songyou Peng, Kyle Genova, ChiyuMaxJiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas A. Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 1, 2, 3, 5, 8

[55] Zekun Qi, Runpei Dong, Guo Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. *arXiv*, abs/2302.02318, 2023. 2, 3

[56] Gordon Guocheng Qian, Xingdi Zhang, Abdullah Hamdi,

and Bernard Ghanem. Pix4point: Image pretrained transformers for 3d point cloud understanding. *arXiv*, abs/2208.12259, 2022. 3

[57] 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 *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021. 1

[58] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 2

[59] Damien Robert, Bruno Vallet, and Loïc Landrieu. Learning multi-view aggregation in the wild for large-scale 3d semantic segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 3

[60] Dávid Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 3, 7

[61] Dávid Rozenberszki, Or Litany, and Angela Dai. Unscene3d: Unsupervised 3d instance segmentation for indoor scenes. *ArXiv*, abs/2303.14541, 2023. 3

[62] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 3, 4

[63] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Eur. Conf. Comput. Vis. (ECCV)*, 2012. 2

[64] Ayca Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *ArXiv*, abs/2306.13631, 2023. 3

[65] Nikolaos Tsagkas, Oisin Mac Aodha, and Chris Xiaoxuan Lu. Vl-fields: Towards language-grounded neural implicit spatial representations. *ArXiv*, abs/2305.12427, 2023. 3

[66] Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. P2p: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 3

[67] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 41:2251–2265, 2017. 6

[68] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas J. Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Eur. Conf. Comput. Vis. (ECCV)*, 2020. 3

[69] Chenfeng Xu, Shijia Yang, Tomer Galanti, Bichen Wu, Xiangyu Yue, Bohan Zhai, Wei Zhan, Péter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Image2point: 3d point-cloud understanding with 2d image pretrained models. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 3

[70] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 2

[71] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and X. Wang. Groupvit: Semantic segmentation emerges from text supervision. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 2, 4

[72] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP: learning unified representation of language, image and point cloud for 3d understanding. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 2

[73] Jihan Yang, Runyu Ding, Zhe Wang, and Xiaojuan Qi. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. *ArXiv*, abs/2304.00962, 2023. 3

[74] Yuan Yao, Yuanhan Zhang, Zhen fei Yin, Jiebo Luo, Wanli Ouyang, and Xiaoshui Huang. 3d point cloud pre-training with knowledge distillation from 2d images. *arXiv*, abs/2212.08974, 2022. 3

[75] Li Yi, Boqing Gong, and Thomas A. Funkhouser. Complete & label: A domain adaptation approach to semantic segmentation of lidar point clouds. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021. 1

[76] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021. 1

[77] Ping Yu, Cheng Sun, and Min Sun. Data efficient 3d learner via knowledge transferred from 2d model. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 3

[78] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 3

[79] Nir Zabari and Yedid Hoshen. Semantic segmentation in-the-wild without seeing any segmentation examples. *arXiv*, abs/2112.03185, 2021. 1

[80] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 2

[81] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 14388–14397, 2021. 2

[82] Karim Abou Zeid, Jonas Schult, Alexander Hermans, and Bastian Leibe. Point2vec for self-supervised representation learning on point clouds. 2023. 3

[83] Yi Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chao 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 *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 2

[84] Junbo Zhang, Guofan Fan, Guanghan Wang, Zhengyuan Su, Kaisheng Ma, and Li Yi. Language-assisted 3d feature learning for semantic scene understanding. In *AAAI Conf. Artif. Intell. (AAAI)*, 2023. 3

[85] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bingyan

Zhao, Dong Lei Wang, Yu Jiao Qiao, and Hongsheng Li. Point-m2ae: Multi-scale masked autoencoders for hierarchical point cloud pre-training. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 3

[86] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Jiao Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021. 2

[87] Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In *CVPR*, 2023. 3

[88] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Int. Conf. Comput. Vis. (ICCV)*, 2021. 3

[89] Zihui Zhang, Bo Yang, Bing Wang, and Bo Li. Growsp: Unsupervised semantic segmentation of 3d point clouds. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 3

[90] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel C. F. Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021. 2

[91] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 1, 2, 3, 5

[92] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krahenbuhl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 2