3D Annotation-Free Learning by Distilling 2D Open-Vocabulary Segmentation Models for Autonomous Driving

Boyi Sun[1, 2], **Yuhang Liu**[1, 2], **Xingxia Wang**[1], **Bin Tian**[1, 3], **Long Chen**[1, 3], **Fei-Yue Wang**[1*]

[1]Institute of Automation, Chinese Academy of Sciences
[2]Zhongke JingYu Sensing Technology Co., Ltd
[3]Waytous
{sunboyi2024, liuyuhang2021, wangxingxia2022, bin.tian, long.chen, feiyue.wang}@ia.ac.cn

Abstract

Point cloud data labeling is considered a time-consuming and expensive task in autonomous driving, whereas annotation-free learning training can avoid it by learning point cloud representations from unannotated data. In this paper, we propose AFOV, a novel 3D **A**nnotation-**F**ree framework assisted by 2D **O**pen-**V**ocabulary segmentation models. It consists of two stages: In the first stage, we innovatively integrate high-quality textual and image features of 2D open-vocabulary models and propose the Tri-Modal contrastive Pre-training (TMP). In the second stage, spatial mapping between point clouds and images is utilized to generate pseudo-labels, enabling cross-modal knowledge distillation. Besides, we introduce the Approximate Flat Interaction (AFI) to address the noise during alignment and label confusion. To validate the superiority of AFOV, extensive experiments are conducted on multiple related datasets. We achieved a record-breaking 47.73% mIoU on the annotation-free 3D segmentation task in nuScenes, surpassing the previous best model by 3.13% mIoU. Meanwhile, the performance of fine-tuning with 1% data on nuScenes and SemanticKITTI reached a remarkable 51.75% mIoU and 48.14% mIoU, outperforming all previous pre-trained models.

Code — https://github.com/sbysbysbys/AFOV

Introduction

As neural-network-based 3D scene perception methods, *e.g.*, object detection (Shi, Wang, and Li 2019; Shi et al. 2020; Zhou and Tuzel 2018; Lang et al. 2019), point cloud segmentation (Qi et al. 2017a,b; Choy, Gwak, and Savarese 2019), *etc.*, become increasingly complex in their network architectures with a growing number of parameters, methods relying solely on enhancing model structures are reaching a point of saturation. Meanwhile, approaches to enhancing model performance through data-driven methods heavily rely on time-consuming and expensive manual annotations. Due to constraints such as insufficient class annotations, applying traditional point cloud perception methods to large-scale unlabeled data meets significant challenges.

Annotation-free learning is a powerful machine learning paradigm that enables representation learning from data without the need for manual supervision. Multi-modal annotation-free algorithms for 3D scene perception tasks integrate the internal structure of point clouds with image or text knowledge to generate objectives. They bypass the costly manual annotations, bridging the gap between traditional 3D perceptual models and unannotated data.

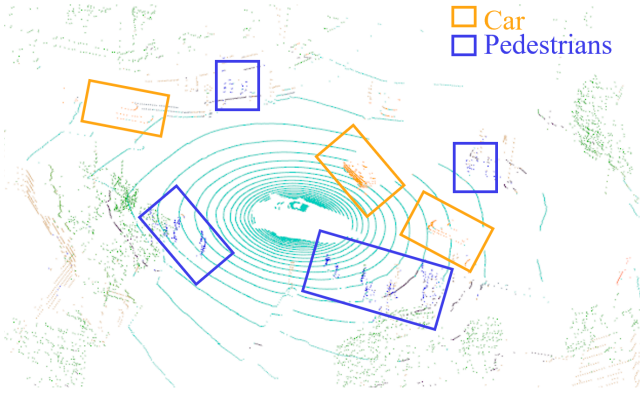

Figure 1: Segmentation results of AFOV annotation-free training. More illustrations are presented in Appendix D.

Existing 3D annotation-free methods (Chen et al. 2023b; Zeng et al. 2023; Chen et al. 2023a; Peng et al. 2023) aim to transfer knowledge from visual foundation models (VFMs), *e.g.*, Contrastive Vision-Language Pre-training (CLIP) (Radford et al. 2021) or Segment Anything (SAM) (Kirillov et al. 2023), to point cloud representations. However, 3D annotation-free models based on CLIP (Radford et al. 2021) suffer from intolerable noise, while SAM (Kirillov et al. 2023) fails to correspond texts and images. Therefore, we seek high-quality image segmentation models with textual correspondences to serve as teacher models for 3D annotation-free learning. Recently, CLIP-based 2D open-vocabulary segmentation models (Zhou, Loy, and Dai 2022; Yu et al. 2023; Xu et al. 2023; Cho et al. 2023) have demonstrated excellent performances. They employ contrastive learning to extract textual and image features from a shared embedding space and are capable of segmenting and identifying objects from a set of open classes in various environments. These models provide us with image segmentation and labels corresponding to the segmented regions,

*Corresponding Author.

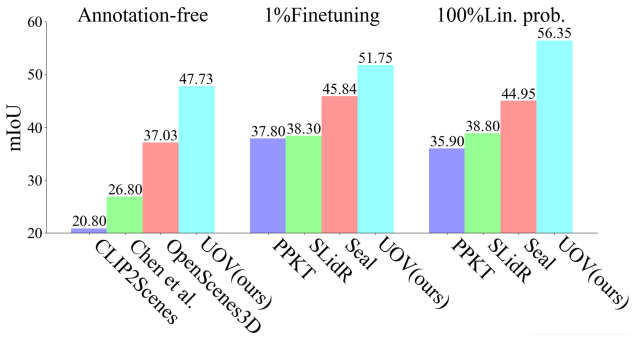

Figure 2: Performance of AFOV on nuScenes.

as well as easy-to-extract text and image representations. Meanwhile, they significantly outperform other Visual Language Models (VLMs) (Radford et al. 2021; Zou et al. 2024, 2023; Li et al. 2023) in open-vocabulary segmentation tasks.

In this paper, we propose AFOV, a 3D **A**nnotation-**F**ree framework by distilling 2D **O**pen-**V**ocabulary segmentation models. It aims to address the difficulties of point cloud perception using unannotated data through 3D annotation-free learning. AFOV adopts a novel two-stage strategy: The first stage uses Tri-Modal contrastive Pre-training (TMP) to warm up the network parameters, where we innovatively incorporate the textual information to enhance the semantic perception of point cloud representations; The second stage is pseudo-label guided annotation-free training, completing the knowledge distillation from 2D to 3D. Furthermore, to address the perceptual limitations of AFOV, such as noise during alignment and label confusion, we introduce Approximate Flat Interaction (AFI). It provides a robust error correction mechanism for the above two stages through point cloud spatial interaction.

In the experiment, We selected four CLIP-based open-vocabulary segmentation models, *ie.*, MaskCLIP (Zhou, Loy, and Dai 2022), FC-CLIP (Yu et al. 2023), SAN (Xu et al. 2023), CAT-Seg (Cho et al. 2023), as teacher models. To validate the performance of our method in annotation-free point cloud segmentation tasks, extensive experiments on multiple autonomous driving datasets were conducted. Firstly, AFOV achieved a remarkable improvement of 3.13% mIoU, reaching a Top-1 accuracy of 47.74% mIoU in benchmark testing of annotation-free 3D segmentation on nuScenes (Caesar et al. 2020). Furthermore, treating AFOV as a pre-trained model, we conducted 1% data fine-tuning and 100% data linear-probing experiments on nuScenes, yielding mIoU scores of 51.75% and 56.35%. Compared to the current best pretraining method, AFOV achieved improvements of 4.16% and 4.81% mIoU, respectively. When fine-tuning with 1% data on SemanticKITTI (pre-training on nuScenes), AFOV achieved a 48.14% mIoU. AFOV demonstrated state-of-the-art performance across various experiments, validating its effectiveness.

Compared to scene understanding work based on point clouds (such as OpenScene), the introduction of text information allows AFOV to generate pseudo labels for knowledge distillation. As a result, AFOV directly predicts outcomes, whereas scene understanding models match features between point clouds and text. As with fully supervised closed-set point cloud segmentation models, the directly predicted output is far superior to the output obtained through feature matching.

Unlike most previous annotation-free 3D segmentation models, all the knowledge for training AFOV comes from state-of-the-art 2D open-vocabulary segmentation models. These open-vocabulary segmentation models (such as FC-CLIP, CAT-Seg, and SAN) perform far better than other VLMs like maskCLIP, SAM, and SEEM. They can also extract masks, labels, and features, avoiding noise accumulation between different backbones. Therefore, AFOV presents significant advantages over previous related works.

In conclusion, the main contributions of this work are:

- We introduce a novel and efficient two-stage annotation-free training framework, AFOV, which comprehensively utilizes state-of-the-art 2D open-vocabulary segmentation models for knowledge distilling.
- AFOV innovatively introduces TMP and AFI, addressing the issues in previous works. Moreover, we introduce the superpixel-superpoint into annotation-free 3D segmentation for the first time.
- Experimentally, our approach not only breaks through in annotation-free semantic segmentation (Fig. 1), but also notably outperforms prior state-of-the-art methods in other downstream tasks (Fig. 2).

Related Work

CLIP-based 2D Open-Vocabulary Segmentation

2D open-vocabulary segmentation models aim to segment all categories in the real world. Traditional open-vocabulary image segmentation models (Zhao et al. 2017; Xian et al. 2019; Bucher et al. 2019) attempt to learn image embeddings aligned with text embeddings. Inspired by Visual Language Models (VLMs), *e.g.*, CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021), which have demonstrated remarkable performance in 2D tasks, recent studies have attempted to transfer CLIP's outstanding zero-shot segmentation capability to open-vocabulary tasks (Xu et al. 2023; Cho et al. 2023; Zhou, Loy, and Dai 2022; Yu et al. 2023; Li et al. 2022a; Ghiasi et al. 2022; Ding et al. 2022; Xu et al. 2022; Liang et al. 2023). In notable works, LSeg (Li et al. 2022a) learns pixel-level visual embeddings from CLIP, marking the first exploration of CLIP's role in language-driven segmentation tasks. More recently, MaskCLIP (Zhou, Loy, and Dai 2022) obtains pixel-level embeddings by modifying the CLIP image encoder; SAN (Xu et al. 2023) augments CLIP with lightweight side networks to predict mask proposals and categories; CAT-Seg (Cho et al. 2023) proposes a cost-aggregation-based method to optimize the image-text similarity map. Additionally; FC-CLIP (Yu et al. 2023) integrates all components into a single-stage framework using a shared frozen convolutional CLIP backbone. These works utilize CLIP as the central component of their network,

granting them robust segmentation and recognition capabilities, as well as an image-text-aligned structure. Consequently, they can provide us with high-quality knowledge.

3D Representation Learning Based on 2D-to-3D Knowledge Distillation

Unsupervised learning can be utilized for learning point cloud representations. Mainstream 3D unsupervised pre-training methods are mainly reconstruction-based (Boulch et al. 2023; Lin and Wang 2022; Hess et al. 2023; Min et al. 2023), or contrast-based (Li et al. 2022b; Xie et al. 2020; Zhang et al. 2021; Liu et al. 2021; Sautier et al. 2022; Mahmoud et al. 2023; Liu et al. 2023). However, many of these methods are constrained by the quantity of point clouds, limiting their applicability to single-object or indoor scene learning. Prior attempts, such as PointContrast (Xie et al. 2020), DepthContrast (Zhang et al. 2021), SegContrast (Nunes et al. 2022), and PPKT (Liu et al. 2021), have built contrastive objectives on large-scale point clouds. Additionally, SLidR (Sautier et al. 2022) adopts a novel approach by leveraging a superpixel-superpoint correspondence for 3D-to-2D spatial alignment, showing promising performance on autonomous driving datasets. Built upon SLidR, SEAL (Liu et al. 2023) employs VLMs to aid in superpixel generation.

Recently, inspired by the achievements of CLIP (Radford et al. 2021), numerous works have focused on reproducing the excellent performance demonstrated by CLIP in 3D annotation-free tasks, not limited to pre-training. In 3D scene understanding, CLIP2Scene (Chen et al. 2023b), similar to OpenScene (Peng et al. 2023) and OV3D (Jiang, Shi, and Schiele 2024), embeds the knowledge of CLIP feature space into representations of 3D point cloud, enabling annotation-free point cloud segmentation; PLA (Ding et al. 2023) and RegionPLC (Yang et al. 2023) accomplish scene understanding through point-language alignment or contrastive learning framework; VLM2Scenes (Liao, Li, and Ye 2024) exploits the potential of VLMs; and CLIP2 (Zeng et al. 2023) demonstrates perfect zero-shot instance segmentation performance through language-3D alignment at the semantic level and image-3D alignment at the instance level. Unlike the others, Chen *et al.* (Chen et al. 2023a) utilizes CLIP to generate pseudo-labels and uses SAM (Kirillov et al. 2023) to assist in denoising.

Method

Extracting Knowledge from CLIP-based 2D Open-Vocabulary Segmentation Models

In perceptual approaches to unknown classes in 2D, unlike zero-shot learning, open-vocabulary learning uses language data as supervision. In terms of network structure, MaskCLIP changes the image encoder of CLIP to propose pixel-level representations instead of image-level. SAN proposes a side adapter network attached to a frozen CLIP encoder; CAT-Seg employs a cost-aggregation-based method to improve CLIP; FC-CLIP adds a decoder, mask generator, in-vocabulary classifier, and out-vocabulary classifier after freezing the CLIP backbone. In most cases, the mask generator operates independently of the class generator. Pixel-level features generated by the modified CLIP-based network are max-pooled for each mask, and the objective loss is computed with the text features.

We can notice that, regardless of whether the CLIP backbone is frozen, whether the CLIP network's architecture is modified, or whether additional network structures are appended to the side or rear of the CLIP network, the essence of these CLIP-based models lies in aligning image features with text features through contrastive learning.

The aforementioned methods hold a similar view with contrastive learning for point cloud pre-training. The difference is that the latter uses image-point cloud contrastive learning (Liu et al. 2021; Sautier et al. 2022) (some of the work uses data augmentation for single-modal contrastive learning). Most of them use SLIC (Achanta et al. 2012), SAM (Kirillov et al. 2023), and SEEM (Zou et al. 2024) to guide mask segmentation and choose ResNet (He et al. 2016) as the image encoder, which means that mask segmentation and mask feature generation are two completely independent modules. This not only increases the training time but also makes noise easily stack up across different models. At the same time, the lack of language guidance during segmentation will lead to a more random mask granularity. Fortunately, 2D open-vocabulary segmentation models perfectly address this issue, as we can not only extract labels and image embeddings from them but also obtain segmentation with appropriate granularity.

To summarize, we extract four interrelated, synchronously generated knowledge from CLIP-based open-vocabulary segmentation models: 1) images' segmentations as masks $M_{\mathcal{I}}$ from image set $\mathcal{I}$; 2) corresponding labels L_M for $M_{\mathcal{I}}$; 3) image features corresponding to $M_{\mathcal{I}}$, denoted as F_M; and 4) text features F_T. Each of the above knowledge will play an important role in the following sections, as shown in Fig. 3.

Baseline of AFOV

Given a point cloud $\mathcal{P} = \{(p_n, e_n)|n = 1, \ldots, N\}$, where $p_n \in \mathbb{R}^3$ represents the 3D coordinates of a point, $e_n \in \mathbb{R}^E$ denotes the point's features. $L = \{l_n|n = 1, \ldots, N\}$ are the labels of $\mathcal{P}$ and $\mathcal{I} = \{i_k|k = 1, \ldots, K\}$ represents the images captured by a synchronized camera at the same moment. In contrast to supervised methods, our task does not utilize labels L during training. We choose to employ a simple way of generating pseudo-labels for point clouds with the assistance of image segmentation: With masks $M_{\mathcal{I}} = \{m_r|r = 1, \ldots, R\}$ obtained from image set $\mathcal{I}$ as described in the above section, we use the labels L_M corresponding to $M_{\mathcal{I}}$ as the pseudo-label $L_{\text{pixel}}^{\text{pseudo}}$ for pixels in every mask $m_r \in M_{\mathcal{I}}$. By leveraging known sensor calibration parameters, we establish a mapping $\Gamma_{\text{camera}\leftarrow\text{LiDAR}}$ to bridge the gap between domains of point clouds and images. Pseudo-labels $L_{\mathcal{P}}^{\text{pseudo}} = \{l_{n_0}^{\text{pseudo}}|n_0 = 1, \ldots, N_0\}$ for point clouds $\mathcal{P}$ are generated through $L_{\text{pixel}}^{\text{pseudo}}$ and mapping $\Gamma_{\text{camera}\leftarrow\text{LiDAR}}$. For a 3D backbone $\mathcal{F}_{\theta_p} : \mathbb{R}^{N\times(3+L)} \rightarrow \mathbb{R}^{N\times D}$ with the learnable parameter θ_p, we train θ_p with pseudo-labels $L_{\mathcal{P}}^{\text{pseudo}}$. Given the sparsity of point clouds, it is obvious that $\Gamma_{\text{camera}\leftarrow\text{LiDAR}}$

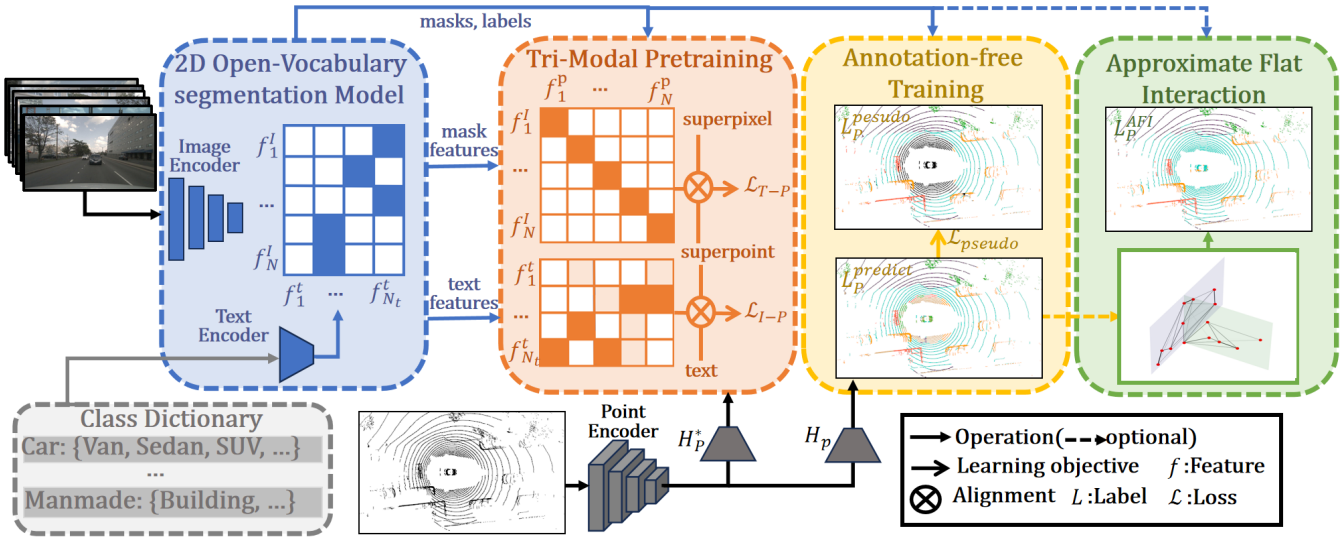

Figure 3: Overview of AFOV, which consists of two stages: Tri-Modal Pre-training (TMP) and Annotation-free training (AFOV-baseline). Both stages leverage masks and mask labels extracted from 2D open-vocabulary segmentation models, while mask features and text features are employed only in TMP. TMP enhances scene understanding through contrastive losses: superpixel-superpoint loss $\mathcal{L}_{I-P}$ and text-superpoint loss $\mathcal{L}_{T-P}$, while our baseline employs pseudo-labels to supervise the 3D network. Additionally, to bridge dataset classes and open vocabularies, we introduce a class dictionary. The Approximate Flat Interaction (AFI) optimizes the results by spatial structural analysis in a broad perception domain.

is not surjective. It is important to note that $\Gamma_{\text{camera}\leftarrow\text{LiDAR}}$ is also not injective, as the projection area of LiDAR is not entirely covered by cameras, resulting in obvious pseudo-label-blank areas in point cloud $\mathcal{P}$. After knowledge distillation, these untrained regions exhibit label confusion, which will be discussed in the next section.

To align open vocabularies with the stuff-classes of autonomous driving datasets, we employ a class dictionary $\mathcal{C} = \{c_i : [t_1^{c_i}, \ldots, t_{n_{c_i}}^{c_i}] | i = 1, \ldots, N^{\mathcal{C}}\}$, where $N^{\mathcal{C}}$ represents the number of stuff-classes. Texts belonging to the same class $t_{j_1}, t_{j_2} \in c_i$ are uniformly mapped to the pseudo-label corresponding to c_i, which implies that points corresponding to t_{j_1} and t_{j_2} are positive samples for each other.

Tri-Modal Contrastive Pre-training (TMP)

In this section, we introduce Tri-Modal contrastive Pre-training (TMP). TMP innovatively integrates textual information into pre-training and removes the 2D backbone through pre-generation of the features, demonstrating excellent performance in both annotation-free training and fine-tuning. The illustration of TMP can refer to Fig. 3.

Synchronous Generation of Knowledge To the best of our knowledge, most existing 3D pre-training methods (Liu et al. 2021; Sautier et al. 2022; Liu et al. 2023) generate masks and mask features asynchronously for 2D-3D contrastive learning, which causes noise aggregation between different backbones. In TMP, we address this issue by synchronously generating masks and features before pre-training.

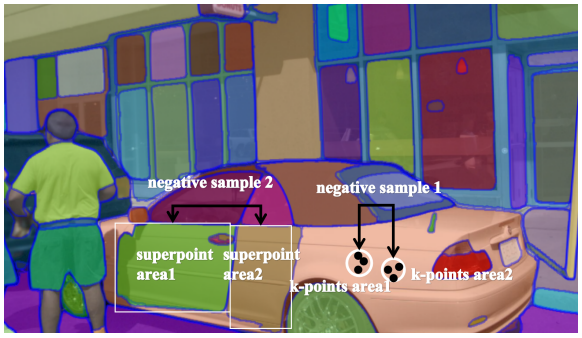

Figure 4: Illustrating two examples of potential "self-conflicts" based on SAM segmentation.

Integrating Textual Information in TMP We aim to incorporate text-3D contrastive learning into pretraining. For different instances of the same category (e.g., Car A and Car B), they share the same textual features but have different image features. Theoretically, 2D-3D contrastive learning provides rich features from the instance level (needed for pretraining); text-3D contrastive learning, on the other hand, offers direct semantic embeddings for the 3D backbone from the semantic level (helpful for annotation-free segmentation). So the design intention of TMP is: Can integrating textual information into pre-training accomplish multiple tasks (pretraining and unannotated segmentation)? Experiment removing text-3D contrastive learning is conducted in Appendix B.

"Self-Conflict" Regions with the same semantics may be divided into multiple parts (in 2D or 3D). During contrastive learning, different parts act as "negative samples" to each other, causing features of parts with the same semantics to be pushed apart in the feature space. This phenomenon is called "self-conflict". The usage of superpixels and superpoints for segmentation in SLidR (Sautier et al. 2022) is inspiring, which partly alleviates the issue of "self-conflict" caused by point-level point set partitioning (*e.g.* k-NN), such as "negative sample 1" in Fig. 4. In Appendix B, we conduct an ablation study on the use of superpixels and superpoints, and the results demonstrate their effectiveness.

However, the "self-conflict" caused by the mask granularity randomness of some VLMs has not been properly addressed, such as "negative sample 2" in Fig. 4. For instance, SAM (Kirillov et al. 2023) might separate the car-door and car-window because SAM is unaware of the appropriate segmentation granularity (both should belong to the same class: "car"). This issue arises due to the lack of textual assistance. When guided by the text, a 2D open-vocabulary segmentation model would treat the entire car as a whole, thereby avoiding "self-conflict".

Superpixel-Superpoint Generation Given the point cloud $\mathcal{P}$ and images $\mathcal{I}$, we have generated masks $M_{\mathcal{I}}$ and use $\Gamma_{\text{camera}\leftarrow\text{LiDAR}}$ to map the labels L_M of masks to $\mathcal{P}$. We regard the set of pixels with corresponding points in the same mask $m_r \in M_{\mathcal{I}}$ as a superpixel $S_r^{\text{pixel}} \in \mathcal{S}^{\text{pixel}}$, while the corresponding region of the point cloud as a superpoint $S_r^{\text{point}} \in \mathcal{S}^{\text{point}}$, $\mathcal{S}^{\text{pixel}}$ and $\mathcal{S}^{\text{point}}$ establish a bijection and ensuring $|\mathcal{S}^{\text{pixel}}| = |\mathcal{S}^{\text{point}}| = R^{\mathcal{S}} \leq |M_{\mathcal{I}}| = R$. Assuming the point cloud backbone $\mathcal{F}_{\theta_p}$ comes with an output head H_p, we replace H_p with a trainable projection head H_p^*, projecting the point cloud feature f_p of $p \in \mathcal{P}$ into a D^*-dimensional space such that $Dim_{f_M} = Dim_{f_T} = D^*$, here $f_M \in F_M$ and $f_T \in F_T$ refer to mask features and text features provided by the 2D open-vocabulary semantic segmentation models. Firstly, we apply average pooling and normalization to each group of pixel features $F_{r0} = \{f_p | p \in S_{r0}^{\text{point}}\}$ guided by superpoints to extract the superpoint embeddings $f_{r0}^{\text{p}} \in F^{\text{superpoint}}$. Then, we normalize F_M as the superpixel embeddings $F^{\text{superpixel}}$ and consider the masks-corresponding text features as F^{text}. Finally, we employ a tri-modal contrastive loss to align $F^{\text{superpixel}} - F^{\text{superpoint}}$, $F^{\text{text}} - F^{\text{superpoint}}$.

Tri-Modal Contrastive Loss Superpixel-guided contrastive learning operates at the object level or semantic level, rather than at the pixel or scene level. The contrastive loss between $F^{\text{superpixel}}$ and $F^{\text{superpoint}}$ is formulated as:

$$\begin{aligned}\mathcal{L}_{I-P} &= \mathcal{L}(F_{\text{superpixel}}, F_{\text{superpoint}}) \\ &= -\frac{1}{R'}\sum_{i=0}^{R'} \log\left[\frac{e^{(\langle f_i^I, f_i^p\rangle/\tau)}}{\sum_{j\neq i} e^{(\langle f_i^I, f_j^p\rangle/\tau)} + e^{(\langle f_i^I, f_i^p\rangle/\tau)}}\right],\end{aligned} \tag{1}$$

where $f_i^I \in F^{\text{superpixel}}$ is the feature of i_{th} superpixel and $f_j^p \in F^{\text{superpoint}}$ is the feature of j_{th} superpoint. $\langle\cdot\rangle$ denotes the cosine similarity and τ denotes the temperature coefficient. R' is the mini-batch size.

Unlike the superpixel-superpoint contrastive loss, the text-superpoint contrastive loss does not exhibit "self-conflict" on classes of a dataset. However, to ensure the uniformity of knowledge in TMP, we retained the class dictionary $\mathcal{C}$ as discussed in **Baseline of AFOV**. In downstream tasks, texts of the same class $t_{j_1}, t_{j_2} \in c_i$ should be considered as positive samples for each other, so treating the point cloud regions corresponding to t_{j_1}, t_{j_2} as "negative samples" will inadvertently cause "self-conflict". Therefore, for text $t_{j_0} \in c_i$, we utilize the text feature's cosine similarity $\langle f_{t_{j_0}}, f_{t_{j_s}}\rangle$ weighted for other texts in the same class $\{t_{j_s} \in c_i, j_s = \{1, \ldots, n_{c_i}\} \neq j_0\}$ as "semi-positive" samples to compute $\mathcal{L}_{T-P}$:

$$\begin{aligned}&\mathcal{L}_{T-P} = \mathcal{L}(F_{\text{text}}, F_{\text{superpoint}}) = \\ &\frac{-1}{R'}\sum_{i=0}^{R'} \log\left[\frac{e^{\left(\frac{\langle f_i^t, f_i^p\rangle}{\tau}\right)}}{\sum_{t_j\neq t_i} e^{\left((1-\alpha_{ij})\frac{\langle f_i^t, f_j^p\rangle}{\tau}\right)} + e^{\left(\frac{\langle f_i^t, f_i^p\rangle}{\tau}\right)}}\right]\end{aligned} \tag{2}$$

$$,\alpha_{ij} = \begin{cases} \langle f_i^t, f_j^t\rangle & t_i, t_j \text{ in same class} \\ 0 & \text{else}\end{cases} \tag{3}$$

where $f_i^t \in F^{\text{text}}$ is the corresponding text feature.

Tri-modal contrastive loss is calculated as:

$$\mathcal{L}_{TMP} = \alpha_{image}\mathcal{L}_{I-P} + \alpha_{text}\mathcal{L}_{T-P}, \tag{4}$$

α_{image}, α_{text} are weights for $\mathcal{L}_{I-P}$ and $\mathcal{L}_{T-P}$.

TMP has the following advantages compared to previous pre-training methods: 1) TMP eliminates the need for image encodings during pre-training. It does not require an image backbone, which reduces the training time. 2) Addition of textual modality. Text-superpoint contrastive learning achieves semantic-level alignment, directly endowing the point cloud backbone with semantic features. 3) Synchronous generation of superpixels and $F^{\text{superpixel}}$. This not only controls segmentation granularity but also prevents the aggregation of noise between different backbones.

Approximate Flat Interaction (AFI)

Through the previous two sections of AFOV, we noticed three technical difficulties for annotation-free semantic segmentation that need to be solved: 1) Unprojected point cloud regions caused by differing or occluded fields of view (FoV) among devices. This directly results in the region of the point cloud outside the image FoV remains untrained for long periods, thus the untrained area suffers from serious label confusion. 2) Label noise in 3D. This arises from matching errors between cameras and LiDAR, as well as noise inherent in 2D open-vocabulary segmentation models. Therefore, we need a robust error correction mechanism for AFOV.

Inspired by Point-NN (Zhang et al. 2023), we propose a non-parametric network for Approximate Flat Interaction (AFI). AFI essentially expects points to interact only among

points lying on approximate planes, thereby preserving labels of relatively small objects, *e.g.*, pedestrians, and vehicles, within a broad perceptual domain. The process of AFI is formulated as:

$$L_{\mathcal{P}}^{AFI} = \text{AFI}(L_{\mathcal{P}}^{predict}, \mathcal{P}, \gamma, (L_{\mathcal{P}}^{pseudo})), \tag{5}$$

$L_{\mathcal{P}}^{predict}$ represents the predictions in **Baseline of AFOV**, and γ indicates the minimum similarity between the directions when their directional features interact. $L_{\mathcal{P}}^{AFI}$, on the other hand, denotes the point cloud labels predicted by the function $AFI(\cdot)$. Meanwhile, we can choose to assist optimization through the pseudo-labels $L_{\mathcal{P}}^{pseudo}$ generated by 2D open-vocabulary segmentation models. A more detailed description of $AFI(\cdot)$ is stated in Appendix A.

During downsampling, AFI passes the directional features of the sampled center point through layer-wise interactions with neighboring points, and binds the correlation between two points based on 1) whether the two points are relevant in the same direction and 2) the tightness of the relevance between correlated directions. Through four rounds of downsampling, point-to-point interactions construct a network that, apart from points at the junctions, AFI ensures the surfaces formed by points on the same network approximate planes, thus tightly controlling interactions among points.

AFI is a robust error correction mechanism for AFOV. The advantages of AFI are evident. 1) Wide-sensing domain: The perception domain for the point clouds with AFI is wide and possesses strong spatial perception capabilities. 2) Detachability: The entire AFI is detachable, and the auxiliary module for 2D images within the AFI is detachable. The effectiveness of AFI can be referenced in Appendix B.

Experiments

Experiments Setup

Datasets To validate the performance of our model, multiple experiments on two large-scale autonomous driving datasets, nuScenes (Caesar et al. 2020) and SemanticKITTI (Behley et al. 2019; Geiger, Lenz, and Urtasun 2012) were conducted, as detailed in **Comparison Results** and **Ablation Study**. In nuScenes, there are 700 scenes for training, while the validation and test set each consist of 150 scenes, comprising a total of 16 semantic segmentation classes. During pre-training, only the train set was utilized, while we validated using specific scenes separated from the train set. SemanticKITTI has 19 classes, with its 22 sequences partitioned into specific train, validation, and test sets.

Implementation Details We followed the training paradigm of SLidR (Sautier et al. 2022), employed MinkowskiNet18 (Choy, Gwak, and Savarese 2019) as the 3D backbone, and used a linear combination of the cross-entropy and the Lovász loss (Berman, Triki, and Blaschko 2018) as training objective in annotation-free and downstream tasks. For 2D open-vocabulary segmentation models, we employed FC-CLIP (Yu et al. 2023), SAN (Xu et al. 2023), CAT-Seg (Cho et al. 2023) for both TMP and annotation-free training, while using MaskCLIP (Zhou, Loy, and Dai 2022) as a control group. The generation of mask features and text features were synchronized with the masks and mask labels. FC-CLIP (Yu et al. 2023) employed panoptic segmentation, distinguishing different instances on thing-classes. MaskCLIP (Zhou, Loy, and Dai 2022), SAN (Xu et al. 2023), and CAT-Seg (Cho et al. 2023) utilized semantic segmentation, not distinguishing instances with the same semantics in both TMP and annotation-free training. Their mask features are selected as the average pool of pixel features in semantically identical regions. In Tri-Modal contrastive Pre-training (TMP), our network was pre-trained for 40 epochs on 4 V100 GPUs with a batch size of 4, which takes about 80 hours. For annotation-free training in **Baseline of AFOV** and other downstream tasks, the network was trained for 5 epochs and 30 epochs on a single V100 GPU, each task taking approximately 3 hours and a batch size of 16. On a 4090 GPU, annotation-free training for 5 epochs only took one hour. The temperature coefficient τ in Eq. 1,2 was set to 0.07, and the optimal results achieved for Eq. 4 when $\alpha_{\text{image}} = \alpha_{\text{text}} = 0.5$. In Eq. 5, the minimum similarity γ between the directions when their directional features interact, was set to 0.995, implying that the maximum angular disparity of two interact point features is about 5.7°. In the network structure of AFI, downsampling was performed four times, with the downsampling rate being $1/3$ for the last three times. Additional details about AFI are provided in Appendix A.

Method	Annotation Ratio	Image Infer	3D backbone	mIoU
CLIP2Scene [CVPR'23]	0%	X	SPVCNN	20.80
Chen et al. [NeurIPS'23]	0%	X	MinkowskiNet	26.80
OpenScene [CVPR'23]	0%	X	MinkowskiNet	41.30
OpenScene-LSeg	0%	X	MinkowskiNet	35.50
OV3D [CVPR'24]	0%	X	MinkowskiNet	44.60
AFOV(ours)+SAN	0%	X	MinkowskiNet	**47.73**
OpenScene [CVPR'23]	0%	✓	MinkowskiNet	36.30
OpenScene-LSeg	0%	✓	MinkowskiNet	42.10
AFOV(ours)+SAN	0%	✓	MinkowskiNet	**47.89**
-	100%		MinkowskiNet	74.66

Table 1: 3D annotation-free semantic segmentation results (% mIoU) on nuScenes (Caesar et al. 2020) *val* set.

Comparison Results

Annotation-free Semantic Segmentation In Tab. 1, we compare AFOV with the most closely related works on 3D semantic segmentation using the unannotated data of nuScenes: CLIP2Scene (Chen et al. 2023b) designs a semantic-driven cross-modal contrastive learning framework; Chen *et al.* (Chen et al. 2023a) utilizes CLIP and SAM for denoising; OpenScene (Peng et al. 2023) extracts 3D dense features from an open-vocabulary embedding space using multi-view fusion and 3D convolution; OV3D (Jiang, Shi, and Schiele 2024) seamlessly aligning 3D point features with entity text features. The optimal result of AFOV's single-modal annotation-free segmentation reaches 47.73% mIoU, surpassing the previous best method by 3.13% mIoU. Under image assistance, it achieves 47.89% mIoU. The gap between AFOV and the fully supervised same backbone is

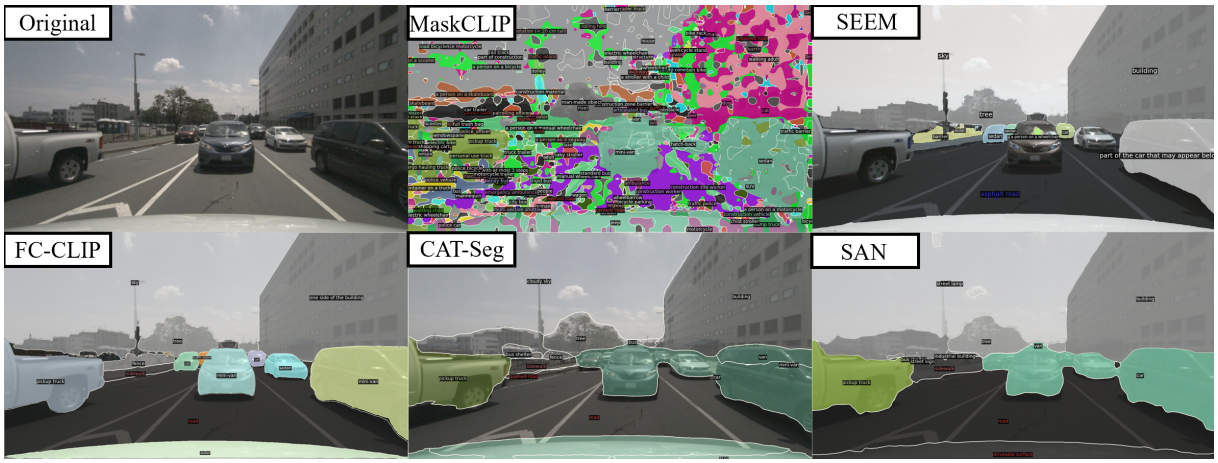

Figure 5: Illustration of image segmentation results of various 2D open-vocabulary segmentation models. We observe that MaskCLIP (pixel-level CLIP) exhibits label confusion and high error rates in semantic segmentation. The output of SEEM not only suffers from missing masks but also contains incorrect mask annotations. More results are provided in Appendix D.

3D Initialization	nuScenes 100%LP	nuScenes 1%Fine-tuning	KITTI 1%Fine-tuning
Random	8.10	30.30	39.50
PointConstrast [ECCV'20]	21.90	32.50	41.10
DepthConstrast [ICCV'21]	22.10	31.70	41.50
PPKT [arXiv'21]	35.90	37.80	44.00
SLidR [CVPR'22]	38.80	38.30	44.60
ST-SLidR [CVPR'23]	40.48	40.75	44.72
Seal [NeurIPS'23]	44.95	45.84	46.63
VLM2Scene [AAAI'24]	51.54	47.59	47.37
ScaLR [CVPR'24]	42.40	40.50	-
AFOV-TMP(ours)+FC-CLIP	44.24	45.73	47.02
AFOV-TMP(ours)+CAT-Seg	43.95	46.61	**48.14**
AFOV-TMP(ours)+SAN	46.29	47.60	47.72
AFOV(ours)+FC-CLIP	52.92	50.58	45.86
AFOV(ours)+CAT-Seg	51.02	49.14	47.59
AFOV(ours)+SAN	**56.35**	**51.75**	46.60

Table 2: Comparisons (% mIoU) of different pre-training methods pre-trained on nuScenes (Caesar et al. 2020) and fine-tuned on nuScenes and SemanticKITTI (Behley et al. 2019). LP denotes linear probing with frozen backbones.

only -26.93% mIoU.

Compared to the multi-task capability of VLMs, state-of-the-art 2D open-vocabulary segmentation models demonstrate greater capability in specialized domains, as shown in the Fig. 5. Selecting professional teacher models enhances the performance of student models effectively.

Comparisons among 3D Pre-training Methods We compared the performance of AFOV-TMP (only employing TMP) and AFOV (employing both steps) against other state-of-the-art methods on multiple downstream tasks in nuScenes and SemanticKITTI (all pre-trained on nuScenes), as shown in Tab. 2. All methods utilized MinkowskiNet as the 3D backbone. Most of the compared state-of-the-art methods utilize point cloud-image contrastive learning. SLidR (Sautier et al. 2022) and ST-SLidR (Mahmoud et al. 2023) employ superpoint-superpixel correspondence granularity; ScaLR (Puy et al. 2024) scales the 2D and 3D backbones and pretraining on diverse datasets; while SEAL (Liu et al. 2023), similar to VLM2Scenes (Liao, Li, and Ye 2024), employs VLMs in distilling. Our approach achieved optimal results with 1% data fine-tuning on nuScenes and SemanticKITTI, reaching 51.75% mIoU and 48.14% mIoU, respectively, demonstrating a respective improvement of +21.45% mIoU and +8.64% mIoU versus random initialization. Compared to the previously best results, AFOV exhibited enhancements of +4.16% mIoU and +0.77% mIoU, respectively. Remarkably, the results of the fully supervised linear probing task on nuScenes reached 56.35% mIoU, displaying an improvement of +4.81% mIoU.

Ablation Study

We conducted a series of ablation experiments on nuScenes. The ablation targets included different teacher models, TMP, AFI, *etc*. The results obtained validate the effectiveness of our designs, especially TMP and AFI. Please refer to Appendix B for details of the ablation study.

Conclusion

We propose AFOV, a versatile two-stage annotation-free framework that serves for both 3D pre-training and annotation-free semantic segmentation, achieving state-of-the-art performance across multiple experiments. The key to AFOV is to leverage the high-quality knowledge of 2D open-vocabulary segmentation models. Moreover, We propose Tri-Modal contrastive Pre-training (TMP) and Approximate Flat Interaction (AFI) for the first time.

We hope that our work will contribute to more in-depth research on 2D-3D transfer learning. Additionally, to the best of our knowledge, there is currently a lack of work on annotation-free training in other 3D perception tasks such as object detection, trajectory tracking, occupancy grid prediction. We expect the emergence of other annotation-free 3D perception approaches.

Acknowledgments

This work is supported by the Beijing Natural Science Foundation (L245025) and the Joint Development of Multi-modal Parallel LiDARs with Waytous Inc.

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *TPAMI*.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*.

Berman, M.; Triki, A. R.; and Blaschko, M. B. 2018. The Lovász-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks. In *CVPR*.

Boulch, A.; Sautier, C.; Michele, B.; Puy, G.; and Marlet, R. 2023. Also: Automotive lidar self-supervision by occupancy estimation. In *CVPR*.

Bucher, M.; Vu, T.-H.; Cord, M.; and Pérez, P. 2019. Zero-shot semantic segmentation. In *NeurIPS*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*.

Chen, R.; Liu, Y.; Kong, L.; Chen, N.; Zhu, X.; Ma, Y.; Liu, T.; and Wang, W. 2023a. Towards label-free scene understanding by vision foundation models. In *NeurIPS*.

Chen, R.; Liu, Y.; Kong, L.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; Qiao, Y.; and Wang, W. 2023b. CLIP2Scene: Towards Label-efficient 3D Scene Understanding by CLIP. In *CVPR*.

Cho, S.; Shin, H.; Hong, S.; An, S.; Lee, S.; Arnab, A.; Seo, P. H.; and Kim, S. 2023. CAT-Seg: Cost Aggregation for Open-Vocabulary Semantic Segmentation. *arXiv preprint arXiv:2303.11797*.

Choy, C.; Gwak, J.; and Savarese, S. 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*.

Ding, J.; Xue, N.; Xia, G.-S.; and Dai, D. 2022. Decoupling zero-shot semantic segmentation. In *CVPR*.

Ding, R.; Yang, J.; Xue, C.; Zhang, W.; Bai, S.; and Qi, X. 2023. PLA: Language-Driven Open-Vocabulary 3D Scene Understanding. In *CVPR*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*.

Ghiasi, G.; Gu, X.; Cui, Y.; and Lin, T.-Y. 2022. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hess, G.; Jaxing, J.; Svensson, E.; Hagerman, D.; Petersson, C.; and Svensson, L. 2023. Masked autoencoder for self-supervised pre-training on lidar point clouds. In *WACV*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

Jiang, L.; Shi, S.; and Schiele, B. 2024. Open-Vocabulary 3D Semantic Segmentation with Foundation Models. In *CVPR*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022a. Language-driven semantic segmentation.

Li, F.; Zhang, H.; Sun, P.; Zou, X.; Liu, S.; Yang, J.; Li, C.; Zhang, L.; and Gao, J. 2023. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*.

Li, Z.; Chen, Z.; Li, A.; Fang, L.; Jiang, Q.; Liu, X.; Jiang, J.; Zhou, B.; and Zhao, H. 2022b. Simipu: Simple 2d image and 3d point cloud unsupervised pre-training for spatial-aware visual representations. In *AAAI*.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*.

Liao, G.; Li, J.; and Ye, X. 2024. VLM2Scene: Self-Supervised Image-Text-LiDAR Learning with Foundation Models for Autonomous Driving Scene Understanding. In *AAAI*.

Lin, Z.; and Wang, Y. 2022. BEV-MAE: Bird's Eye View Masked Autoencoders for Outdoor Point Cloud Pre-training. *arXiv preprint arXiv:2212.05758*.

Liu, Y.; Kong, L.; Cen, J.; Chen, R.; Zhang, W.; Pan, L.; Chen, K.; and Liu, Z. 2023. Segment Any Point Cloud Sequences by Distilling Vision Foundation Models. In *NeurIPS*.

Liu, Y.-C.; Huang, Y.-K.; Chiang, H.-Y.; Su, H.-T.; Liu, Z.-Y.; Chen, C.-T.; Tseng, C.-Y.; and Hsu, W. H. 2021. Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining. *arXiv preprint arXiv:2104.04687*.

Mahmoud, A.; Hu, J. S.; Kuai, T.; Harakeh, A.; Paull, L.; and Waslander, S. L. 2023. Self-Supervised Image-to-Point Distillation via Semantically Tolerant Contrastive Loss. In *CVPR*.

Min, C.; Xiao, L.; Zhao, D.; Nie, Y.; and Dai, B. 2023. Occupancy-mae: Self-supervised pre-training large-scale lidar point clouds with masked occupancy autoencoders. *TIV*.

Nunes, L.; Marcuzzi, R.; Chen, X.; Behley, J.; and Stachniss, C. 2022. SegContrast: 3D point cloud feature representation learning through self-supervised segment discrimination. *RAL*.

Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T.; et al. 2023. Openscene: 3d scene understanding with open vocabularies. In *CVPR*.

Puy, G.; Gidaris, S.; Boulch, A.; Siméoni, O.; Sautier, C.; Pérez, P.; Bursuc, A.; and Marlet, R. 2024. Three pillars improving vision foundation model distillation for lidar. In *CVPR*.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Sautier, C.; Puy, G.; Gidaris, S.; Boulch, A.; Bursuc, A.; and Marlet, R. 2022. Image-to-lidar self-supervised distillation for autonomous driving data. In *CVPR*.

Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; and Li, H. 2020. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *CVPR*.

Shi, S.; Wang, X.; and Li, H. 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*.

Xian, Y.; Choudhury, S.; He, Y.; Schiele, B.; and Akata, Z. 2019. Semantic projection network for zero-and few-label semantic segmentation. In *CVPR*.

Xie, S.; Gu, J.; Guo, D.; Qi, C. R.; Guibas, L.; and Litany, O. 2020. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *ECCV*.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*.

Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; and Bai, X. 2022. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *ECCV*.

Yang, J.; Ding, R.; Wang, Z.; and Qi, X. 2023. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. *arXiv preprint arXiv:2304.00962*.

Yu, Q.; He, J.; Deng, X.; Shen, X.; and Chen, L.-C. 2023. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *arXiv preprint arXiv:2308.02487*.

Zeng, Y.; Jiang, C.; Mao, J.; Han, J.; Ye, C.; Huang, Q.; Yeung, D.-Y.; Yang, Z.; Liang, X.; and Xu, H. 2023. CLIP2: Contrastive Language-Image-Point Pretraining from Real-World Point Cloud Data. In *CVPR*.

Zhang, R.; Wang, L.; Wang, Y.; Gao, P.; Li, H.; and Shi, J. 2023. Parameter is not all you need: Starting from non-parametric networks for 3d point cloud analysis. *arXiv preprint arXiv:2303.08134*.

Zhang, Z.; Girdhar, R.; Joulin, A.; and Misra, I. 2021. Self-supervised pretraining of 3d features on any point-cloud. In *ICCV*.

Zhao, H.; Puig, X.; Zhou, B.; Fidler, S.; and Torralba, A. 2017. Open vocabulary scene parsing. In *ICCV*.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract free dense labels from clip. In *ECCV*.

Zhou, Y.; and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*.

Zou, X.; Dou, Z.-Y.; Yang, J.; Gan, Z.; Li, L.; Li, C.; Dai, X.; Behl, H.; Wang, J.; Yuan, L.; et al. 2023. Generalized decoding for pixel, image, and language. In *CVPR*.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2024. Segment everything everywhere all at once. In *NeurIPS*.