RegionPLC: Regional Point-Language Contrastive Learning for Open-World 3D Scene Understanding

Jihan Yang[1*] Runyu Ding[1*] Weipeng Deng[1] Zhe Wang[2] Xiaojuan Qi[1]
[1]The University of Hong Kong [2]SenseTime Research
https://jihanyang.github.io/projects/RegionPLC

Abstract

*We propose a lightweight and scalable **Regional Point-Language Contrastive** learning framework, namely **RegionPLC**, for open-world 3D scene understanding, aiming to identify and recognize open-set objects and categories. Specifically, based on our empirical studies, we introduce a 3D-aware SFusion strategy that fuses 3D vision-language pairs derived from multiple 2D foundation models, yielding high-quality, dense region-level language descriptions without human 3D annotations. Subsequently, we devise a region-aware point-discriminative contrastive learning objective to enable robust and effective 3D learning from dense regional language supervision. We carry out extensive experiments on ScanNet, ScanNet200, and nuScenes datasets, and our model outperforms prior 3D open-world scene understanding approaches by an average of 17.2% and 9.1% for semantic and instance segmentation, respectively, while maintaining greater scalability and lower resource demands. Furthermore, our method has the flexibility to be effortlessly integrated with language models to enable open-ended grounded 3D reasoning without extra task-specific training. Code will be released at github.*

1. Introduction

Open-world 3D scene understanding aims to equip models with the ability to accurately perceive and identify open-set objects and categories from 3D data, such as point clouds. This ability is crucial for real-world applications where objects from open-set categories are prevalent [2, 39]. However, this task poses significant challenges due to the scarcity of dense 3D semantic annotations, which are difficult to gather and scale to a large vocabulary space.

Fortunately, the abundance of paired image and text data from the Internet, featuring a vast semantic vocabulary, has enabled 2D vision-language models to exhibit exceptional open-world image comprehension capabilities. These abilities span various tasks, such as image captioning [1, 31], grounding [24, 32], and dense semantic prediction [19, 20, 44]. Consequently, recent research has been inspired to leverage these models to generate pseudo supervision such as dense semantic features [23, 40] and language descriptions [7, 8] for training 3D models, thereby enabling open-world inference without relying on image modalities.

Despite advancements, existing solutions still exhibit limitations. For instance, feature distillation-based methods [23, 40]– despite harvesting dense supervision– suffer from the constraints of 2D feature qualities and require resource-intensive feature extraction, fusion, and storage processes, preventing them from being scaled up with more advanced 3D architectures and larger 3D datasets. Additionally, while [7] utilizes pseudo 3D-language pairs to enable direct learning from large-vocabulary language supervision, it suffers from sparse supervision provided by image captioning models. Considering the recent success of 2D foundation models in image- and region-level vision-language learning, we explore combining their strengths to enrich vocabulary and construct high-quality region-level 3D-language associations. By doing so, our method can yield denser 3D-language supervision and circumvent the knowledge limitations of a single foundation model, facilitating resource-efficient and large-vocabulary 3D learning.

To this end, we propose a holistic **Region**al **P**oint **L**anguage **C**ontrastive learning framework, named **RegionPLC**. This framework generates and fuses diverse region-level captions from powerful 2D vision-language models, which are subsequently mapped to 3D for constructing region-level 3D and language pairs. These paired data are then incorporated into a region-aware point-discriminative contrastive learning framework, enabling 3D open-world learning from dense language supervision.

Specifically, we begin by conducting a comprehensive examination of various 2D foundation models (e.g., image captioning [31], dense captioning [24, 32], and detection models [44]) along with visual prompting techniques for their capability to generate region-level 3D-language pairs. Based on our examination, we propose a supplementary-oriented fusion strategy that leverages the geometric relationship of regions in 3D space to alleviate ambiguities and

conflicts encountered when combining paired 3D-language data from multiple 2D models, ultimately delivering high-quality dense region-level 3D-language pairs. Furthermore, with region-level language data, we introduce a region-aware point-discriminative contrastive loss that prevents the optimization of point-wise embeddings from being disturbed by nearby points from unrelated semantic categories, enhancing the discriminativeness of learned point-wise embeddings. The region-aware design further normalizes the contribution of multiple region-level 3D-language pairs, regardless of their region sizes, making feature learning more robust. Finally, by harvesting the 3D-language associations, our RegionPLC can be effortlessly integrated with language models to enable open-ended 3D reasoning with grounding abilities without requiring task-specific data for training.

We conduct extensive experiments on ScanNet [6], ScanNet200 [26], and nuScenes [3] datasets, covering both 3D indoor and outdoor scenarios. Our method significantly outperforms existing open-world scene understanding methods, achieving an average of 17.2% gains in terms of unseen category mIoU for semantic segmentation and an average of 9.1% gains in terms of unseen category mAP_{50} for instance segmentation. RegionPLC demonstrates promising zero-shot segmentation performance, attaining 40.5% and 1.8% higher foreground mIoU compared to PLA [7] and OpenScene [23], respectively. Notably, it achieves this performance while consuming only 17% of OpenScene's [23] training cost and 5% of its storage requirements. Furthermore, RegionPLC can also be combined with OpenScene to deliver 5.8% and 10.0% gains in foreground mIoU and mAcc, respectively.

2. Related Work

3D Scene Understanding. 3D semantic and instance segmentation are two fundamental tasks for scene understanding, which predict each point's semantic meaning (and instance IDs) in a 3D point cloud. For semantic feature extraction and prediction, existing approaches design customized point convolutions applied on raw point clouds [28, 33, 35] or employ sparse convolution [10] to develop voxel-based networks [5, 11] or transformers [18] based on 3D grids. For instance-level prediction, representative approaches often use a bottom-up strategy that groups points to form object proposals [16, 29, 30], or first predicts 3D bounding boxes and then refines the object masks using a top-down solution [17, 36, 38]. Though achieving outstanding results on close-set benchmark datasets, they always struggle with open-world recognition.

Open-world 3D Understanding. Open-world 3D understanding [7, 21, 23, 41] aims to recognize novel categories that are unseen during training. Most recently, the high open-world capability of 2D foundation models [1, 25] trained on massive multi-modality data has inspired recent approaches to leverage them for 3D open-world understanding. One line of work [13–15, 23, 27, 42] focuses on incorporating these 2D foundation models in the *inference stage for open-world recognition*, which mainly conducts open-world semantic prediction on the image modality using vision-language models [19, 25, 37] and fuses 2D prediction results into 3D if required. Though promising, they suffer from significant computation and storage overheads during inference and can be sub-optimal to address 3D understanding without learning from 3D geometries.

This paper focuses on another open-world research direction that concentrates on *open-world point cloud learning*. It requires training 3D backbones to enable their open-world capabilities without the image modality dependence during inference and thus have more applicability potential. Along this line of research, some [23, 40, 41] have attempted to distill 2D dense features [19, 37] into 3D backbones for 3D feature learning. However, they still incur high training costs and might inherit 2D prediction failure modes. In addition, Ding *et al*. [7, 8] obtain point-language paired data through image captioning by VL foundation models for training 3D backbones. These methods are scalable toward a large vocabulary space and can be easily integrated with advanced 3D backbones. Despite the advantages, they still suffer from the coarse language supervision.

3. RegionPLC

3.1. Overview

We focus on 3D open-world scene understanding at both semantic and instance levels. During training, given a point cloud of a scene $\mathcal{P} = \{\mathbf{p}\}$, the model can utilize human annotations $\mathcal{Y}$ for base categories $\mathcal{C}^B$, but cannot access annotations for novel categories $\mathcal{C}^N$. During the inference phase, the trained model needs to classify and localize points associated with both base and novel categories ($\mathcal{C}^B \cup \mathcal{C}^N$).

To achieve open-world understanding, apart from the common 3D encoder F_{3D}, we follow [7] to replace the classification layer weights with category embeddings $\mathbf{f}^l$ extracted from a pretrained text encoder F_{text} of CLIP [25] (See Figure 1: Upper). Hence, the prediction process is shown as follows:

$$\mathbf{f}^p = \text{F}_\theta(\text{F}_{\text{3D}}(\mathbf{p})),\ \mathbf{s} = \sigma(\mathbf{f}^l \cdot \mathbf{f}^p),\ \mathbf{o} = \text{F}_{\text{loc}}(\text{F}_{\text{3D}}(\mathbf{p}), \mathbf{s}), \quad (1)$$

where F_θ is the vision-language (VL) adapter to align the feature dimension of the 3D point-wise features $\mathbf{f}^p$ and category embeddings $\mathbf{f}^l$, $\mathbf{s}$ is the semantic classification score, σ is the softmax function, $\mathbf{o}$ is the instance proposal output, and F_{loc} is the localization network [30] for instance segmentation. With these modifications, the model can predict any desired categories by computing similarity between point-wise features and queried category embeddings for open-world inference.

The goal of our RegionPLC is to train such an open-world 3D backbone via dense region-level 3D-language supervision leveraging powerful and diverse 2D foundation

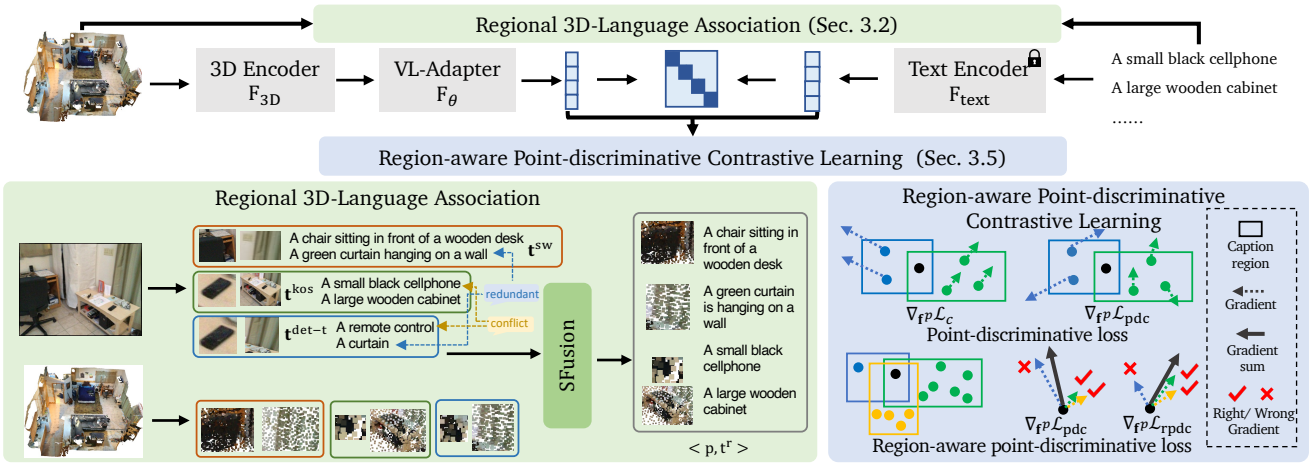


Figure 1. Overview of our regional point-language contrastive learning framework. For regional 3D-language association, We develop a 3D-aware SFusion strategy effectively combining 3D vision-language pairs obtained from multiple 2D foundation models (refer to Sec. 3.2). Upon these 3D-language data, we propose region-aware point-discriminative contrastive learning to facilitate more distinctive and robust representation learning (detailed in Sec. 3.5). Different point & box colors in the bottom-right indicate various 3D-caption pairs.

models, as shown in Figure 1. we first obtain region-level 2D-language pairs through three streams of 2D VL models (*i.e.* image captioning [31], object detection [44] and dense captioning [24, 32]) and then associate them to 3D points (see Sec. 3.2). Then, we comprehensively benchmark and examine these 3D-language pairs from different sources, deriving their merits and shortcomings for 3D learning (see Sec. 3.3). Based on our study, we propose a simple **S**upplementary-oriented **Fusion** (**SFusion**) strategy leveraging their 3D relationships to alleviate redundancies and conflicts in Sec. 3.4, obtaining vocabulary-enriched and denser region-level 3D-language paired data. Finally, upon the 3D-language data, we design a region-aware point-discriminative contrastive learning objective to replace CLIP-style loss for more robust and discriminative feature learning from language supervisions in Sec. 3.5.

3.2. Regional 3D-Language Association from 2D Foundation Models

Here, we first introduce three streams of methods along with two types of visual prompts to extract regional language descriptions from 2D vision-language foundation models: i) object detector with language template; ii) explicit visual prompted image captioning; iii) dense captioning.

Object Detector with language template $\mathcal{G}^{\text{det}}$. The most straightforward manner to obtain regional language supervision is to leverage the category prediction from 2D object detector Detic [44] and then fill the category into a language template as CLIP [25], as illustrated in Figure 2. Thanks to the multi-scale training strategy, object detectors can capture remote and small objects. We denote such regional captions as $\mathbf{t}^{\text{det-t}}$.

Explicit visual prompted image captioning $\mathcal{G}^{\text{prompt}}$. Another intuitive paradigm is first to generate explicit visual

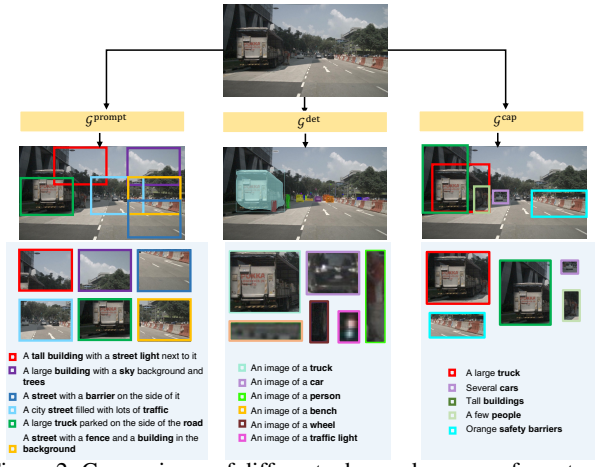


Figure 2. Comparisons of different advanced manners for extracting regional language descriptions with 2D foundation models.

prompts such as boxes and then caption these image patches via image captioning model OFA [31] (refer to Figure 2). As for obtaining explicit visual prompts, we attempt two types: sliding windows and object proposals. Sliding-window-cropped image patches cover all potential semantic regions without being constrained by pre-defined vocabulary space, benefiting open-world tasks but sacrificing precise localization. In contrast, 2D object proposals from detectors provide more accurate object localization but suffer from the limited vocabulary space with pre-defined label space such as LVIS [12]. The obtained dense region-level captions through sliding-window prompts and detector prompts are denoted as $\mathbf{t}^{\text{sw}}$ and $\mathbf{t}^{\text{det-c}}$, respectively.

Dense Captioning $\mathcal{G}^{\text{cap}}$. Apart from the powerful image detectors and image captioning models, recent advances in dense captions and grounding models such as GRiT [32] and Kosmos-2 [24] are trained on the large-scale 2D box and box description pairs. As shown in Figure 2, dense captioners offer precise object localization and rich vocabulary

Method	ScanNet B12/N7		ScanNet annotation-free	
	25K	125K	25K	125K
$\mathbf{t}^{\text{det-t}}$	63.4 / **70.3** / 57.7	**67.4** / 70.5 / **64.6**	37.7 (59.7)	41.9 (64.5)
$\mathbf{t}^{\text{sw}}$	**65.9** / 70.2 / **62.1**	66.0 / 70.2 / 62.3	48.1 (69.2)	47.8 (69.2)
$\mathbf{t}^{\text{det-c}}$	64.4 / 69.9 / 59.7	65.6 / 70.7 / 61.2	41.4 (64.1)	43.8 (65.6)
$\mathbf{t}^{\text{grit}}$	62.7 / **70.3** / 56.6	64.9 / **70.8** / 59.9	**50.5** (**72.2**)	51.3 (**74.2**)
$\mathbf{t}^{\text{kos}}$	64.4 / **70.3** / 59.4	64.6 / 69.8 / 60.2	50.1 (70.7)	**51.7** (72.7)
$\mathbf{t}^{\text{kos}} \cup \mathbf{t}^{\text{det-t}}$	64.6 / 69.8 / 60.2	**67.0** / 69.9 / **64.4**	44.4 (66.3)	**54.0** (**74.5**)
$\mathbf{t}^{\text{kos}} \cup \mathbf{t}^{\text{sw}}$	**65.9** / 69.9 / **62.4**	65.6 / **70.9** / 61.0	**53.5** (72.9)	53.1 (73.9)
$\mathcal{L}_{\mathbf{t}^{\text{kos}}} + \mathcal{L}_{\mathbf{t}^{\text{det-t}}}$	65.0 / **70.2** / 60.5	64.4 / 69.1 / 60.2	51.3 (70.7)	51.2 (72.4)
$\mathcal{L}_{\mathbf{t}^{\text{kos}}} + \mathcal{L}_{\mathbf{t}^{\text{sw}}}$	65.4 / 70.0 / 61.4	64.6 / 70.2 / 59.8	52.9 (**73.6**)	52.7 (73.8)

Table 1. Results of regional caption fusion on base-annotated (hIoU / mIoU$^{\mathcal{B}}$ / mIoU$^{\mathcal{N}}$) and annotation-free (mIoU† (mAcc†), tested on foreground classes only) 3D ScanNet semantic segmentation. $\mathbf{t}^{\text{kos}} \cup \mathbf{t}^{\text{det-t}}$ and $\mathcal{L}_{\mathbf{t}^{\text{kos}}} + \mathcal{L}_{\mathbf{t}^{\text{det-t}}}$ indicate data-level and multi-loss fusion, respectively. Best results are presented in **bold**.

spaces but tend to focus on only salient objects and ignore small and distant objects. We denote captions generated through GRiT [32], Kosmos-2 [24] and Detic [44] with a caption template as $\mathbf{t}^{\text{grit}}$, and $\mathbf{t}^{\text{kos}}$ and $\mathbf{t}^{\text{det-t}}$, respectively.

Associate Points to Dense Captions. Upon above 5 types of regional captions $\mathbf{t}^r = \{\mathbf{t}^{\text{sw}}, \mathbf{t}^{\text{det-c}}, \mathbf{t}^{\text{det-t}}, \mathbf{t}^{\text{grit}}, \mathbf{t}^{\text{kos}}\}$, we associate them to partial point sets through 3D geometry, similar to [7, 23], to pair points and language. Specifically, we begin by projecting the 3D scenes onto 2D images to align points with pixels. Then by connecting the points $\hat{\mathbf{p}}$ within each 2D region to their respective captions, we obtain the regional 3D-language pairs $\langle \hat{\mathbf{p}}, \mathbf{t}^r \rangle$.

3.3. Benchmark and Analysis on Regional 3D-Language Pairs

With the constructed five types of regional 3D-language pairs $\mathbf{t}^r = \{\mathbf{t}^{\text{sw}}, \mathbf{t}^{\text{det-c}}, \mathbf{t}^{\text{det-t}}, \mathbf{t}^{\text{grit}}, \mathbf{t}^{\text{kos}}\}$, the follow-up question is which delivers the best performance on learning 3D open-world representation and how to combine them to obtain enriched vocabulary space and denser regional 3D-language association. Hence, we benchmark them on ScanNet [6] semantic segmentation tasks with different novel categories and 2D image quantities (25K *vs*. 125K). Our benchmark encompasses two settings: i) the B12/N7 setting including 12 annotated base categories and 7 unannotated novel categories, which requires a strong comprehension of a large vocabulary corpus; ii) the annotation-free setting, wherein all categories are novel ones, and thus necessitates both open-vocabulary recognition and precise object localization with only sparse 3D-language pairs.

Complementary cues. As shown in the upper of Table 1, no single type of 3D-language source consistently outperforms others in all settings, and each association has its own merits. For example, $\mathbf{t}^{\text{det-t}}$ inherits the advanced small object localization capabilities (refer to Figure 2 middle for "traffic light" and "wheel" descriptions.), excelling others in the ScanNet B12/N7 (125K). However, it suffers from the limited pre-defined vocabulary space and obtains the worst performance in the annotation-free setting with 17 novel categories. In contrast, dense captioners $\mathbf{t}^{\text{kos}}$ and $\mathbf{t}^{\text{grit}}$ offer salient object localization with semantic-rich vocabulary (refer to Figure 2 right for attribute descriptions), exhibiting superior results on the annotation-free setting. This suggests that different VL models and visual prompts offer various merits and might complement each other.

End-to-end manners scale better. When comparing the performance of utilizing 25K and 125K images, we find that end-to-end trained dense captioners and detectors (*i.e*. $\mathbf{t}^{\text{kos}}$, $\mathbf{t}^{\text{grit}}$ and $\mathbf{t}^{\text{det-t}}$) scale better than the two-stage image captioning manners with visual prompts. The reason might be that end-to-end trained dense caption sources are more consistent on different views and thus yield fewer semantic conflicts when scaling up to more views.

Common combinations are not always effective. As abovementioned, different 3D-language pairs can offer complementary cues. Hence, we examine their synergy effect for better performance. As shown in the bottom of Table 1, we attempt to combine the representatives from three streams of regional caption generation manners $\mathbf{t}^{\text{kos}}$, $\mathbf{t}^{\text{det-t}}$ and $\mathbf{t}^{\text{sw}}$ via data-level and multi-loss fusion. Nevertheless, the performance lift across different settings is not consistent or only shows incremental increases, which suggests the need for a more dedicated fusion strategy to accommodate extensive dense language supervision from multiple sources.

3.4. Boost Synergy of Diverse 3D-language Sources

Motivated by the observations of complementary merits of individual 3D-language sources and their unsatisfactory synergy results, we further study how to combine these varied 3D-language sources effectively and efficiently. In this regard, we propose a Supplementary-orientated Fusion (SFusion) strategy to integrate the most diverse semantic clues while filtering out potential conflicts from different caption sources. As data-level mixing delivers better performance than loss-level combination, we focus on tackling the bottleneck of data-level 3D-language pairs fusion here. When training 3D models on data-level mixed 3D-language pairs, they are learning from a more informative language description, but suffer from sub-optimal performance. This suggests that the main challenges in straightforward data-level mixing are the redundancy and conflicts from different caption sources, especially for highly overlapped point cloud regions (see Figure 1). For those highly overlapped 3D regions with multiple language sources, mutually conflicting descriptions will confuse models, and the overabundance of repetitive language descriptions tends to overwhelm optimization toward easily identifiable areas, leading to sub-optimal performance.

Hence, our SFusion addresses these problems by fusing 3D-language pairs with low 3D overlaps to alleviate potential conflicts in overlapped areas and obtain spatially supplementary 3D-language pairs. Specifically, we first select the most reliable caption source that performs best as the primary 3D-language source $\mathbf{t}^{\text{pri}}$. Then, we compute the

overlap ratio τ of point sets between the primary source and candidate caption sources $\mathbf{t}^{\text{can}}$ on i-th 3D scene as follows,

$$\tau_{jk} = \text{overlap}(\hat{\mathbf{p}}_{ij}^{\text{pri}}, \hat{\mathbf{p}}_{ik}^{\text{can}}), \quad \hat{\tau}_k = \max_j \tau_{jk}, \tag{2}$$

where overlap measures the intersection over union (IoU) between two point sets, $\hat{\mathbf{p}}_{ij}^{\text{pri}}$ and $\hat{\mathbf{p}}_{ik}^{\text{can}}$ are the j-th and k-th point set in the i-th scene from the primary source and candidate source, respectively. Then, we define thresholds T_l and T_h to filter out 3D-language pairs with high overlap ratios from $\mathbf{t}^{\text{can}}$, which might result in redundant or conflict supervision to the primary source. Hence, only candidate 3D-language pairs $\langle \hat{\mathbf{p}}_{ik}^{\text{can}}, \mathbf{t}_{ik}^{\text{can}} \rangle$ with $T_l < \hat{\tau}_k < T_h$ (T_l set to zero) are fused with the primary source. This procedure can be iteratively applied across all candidate caption sources to obtain a collection of 3D-language pairs with low geometrical overlaps. This refined set will serve as the supervision for the follow-up contrastive training. Notice that we also introduce a hyper-parameter $\epsilon \in [0, 1]$ to control the ratio of the primary source and candidate source during fusion, as maintaining the majority of primary sources is beneficial during training with multi-source 3D-language pairs. Our experimental results in Table 7 verify our above claims and demonstrate that our SFusion strategy can significantly boost the combination of multiple language sources.

3.5. Region-aware Point-discriminative Contrastive Learning

After obtaining 3D-language pairs $\langle \hat{\mathbf{p}}, \mathbf{t} \rangle$ for supervision, we proceed to train $\mathbf{F}_{\text{3D}}$ and $\mathbf{F}_\theta$ to align 3D features with language features for open-world learning. We introduce region-aware point-discriminative contrastive loss as below.

CLIP-style Contrastive Loss. We can pull paired 3D features and language features closer while pushing away the unmatched ones through CLIP-style [25] contrastive loss (refer to Figure 1 top right). It can be formulated as:

$$\mathbf{f}^{\hat{p}} = \text{Pool}(\hat{\mathbf{p}}, \mathbf{f}^p), \quad \hat{\mathbf{z}} = \mathbf{f}^{\hat{p}} \cdot \mathbf{F}^t, \quad \hat{\mathbf{s}} = \sigma(\hat{\mathbf{z}}), \tag{3}$$

$$\mathcal{L}_c = -\mathbf{y}^t \cdot \ln \hat{\mathbf{s}}, \tag{4}$$

where $\mathbf{f}^{\hat{p}}$ is the average-pooled region feature, $\text{Pool}(\hat{\mathbf{p}}, \mathbf{f}^{\hat{p}})$ is our custom CUDA operator to gather features $\mathbf{f}^{\hat{p}}$ over point set $\hat{\mathbf{p}}$, $\mathbf{F}^t = [\mathbf{f}_1^t, \mathbf{f}_2^t, \cdots \mathbf{f}_{n_t}^t]$ concatenates all caption embeddings in a scene, $\hat{\mathbf{z}}$ and $\hat{\mathbf{s}}$ measure the similarity and the score probability between a 3D region and all captions, σ is sigmoid function and $\mathbf{y}^t$ is the one-hot label highlighting the position paired with $\hat{\mathbf{p}}$. While CLIP [25] targets learning a global image-level feature for the classification task, it neglects the demand of learning point-wise discriminative features for dense prediction tasks. As shown in Figure 1, the pooling operation will average the point-wise features and make all points in the same region $\hat{\mathbf{p}}$ optimized in the same direction, preventing the learning of discriminative representations for dense prediction tasks. We present more analysis on this undesired effect in the suppl..

Point-discriminative Contrastive Loss. Considering the limitation of CLIP-style loss, we propose a point-discriminative contrastive loss $\mathcal{L}_{\text{pdc}}$ to make the learning of point embedding discriminative. Specifically, for each regional 3D-language pair, instead of aggregating point features into an averaged region-level feature, our $\mathcal{L}_{\text{pdc}}$ directly computes the similarity between point-wise embeddings and caption embeddings. We then pool the logarithm of predicted point-wise probability within $\hat{\mathbf{p}}$ to compute the cross-entropy loss regarding one-hot label $\mathbf{y}^t$ as follows,

$$\mathbf{z} = \mathbf{f}^p \cdot \mathbf{F}^t, \quad \mathbf{s} = \sigma(\mathbf{z}), \quad \mathcal{L}_{\text{pdc}} = -\mathbf{y}^t \cdot \text{Pool}(\hat{\mathbf{p}}, \ln \mathbf{s}), \tag{5}$$

where $\mathbf{z}$ and $\mathbf{s}$ indicate the similarity and probability matrix between point-wise features and all caption embeddings. By doing so, the optimization direction of each point will be adapted to its own point embeddings and thus make them discriminative (refer to Figure 1). More details are included in the supplementary materials.

Region-aware Normalization. Though discriminative, the $\mathcal{L}_{\text{pdc}}$ will back-propagate smaller gradients to points in large regions due to the pooling operation, leading to an implicit bias towards region size which can be harmful to representation learning. To alleviate this issue, we propose a region-aware factor to normalize $\mathcal{L}_{\text{pdc}}$ by the region size, to ensure an equivalent gradient scale on points in each region regardless of its size.Obtained region-aware loss $\mathcal{L}_{\text{rpdc}}$ as follows,

$$\mathcal{L}_{\text{rpdc}} = -\alpha_r \mathcal{L}_{\text{pdc}}, \quad \alpha_r = \frac{n_t \cdot \text{card}(\hat{\mathbf{p}})}{\sum_i^{n_t} \text{card}(\hat{\mathbf{p}}_i)}, \tag{6}$$

where α_r is the region-aware normalization factor, n_t is the number of 3D-language pairs each scene.

Analysis. With point-discriminative and region-aware properties, our $\mathcal{L}_{\text{rpdc}}$ facilitates more superior and robust representation learning. It allows each point to grasp its unique semantics without disruptions from other unrelated points (refer to Figure 1 right). This is especially vital for annotation-free dense prediction to segment object boundaries without any annotation (see Table 6 for verification). Moreover, the region-aware factor in $\mathcal{L}_{\text{rpdc}}$ provides a more robust optimization procedure. As depicted in the right section of Figure 1, points associated with multiple captions are normalized to a similar gradient scale. When multiple captions reach a consensus, this leads to consistent gradient directions, thereby encouraging them to be optimized in a unified direction. Conversely, when multiple captions conflict, this leads to inconsistent gradient directions and thus discourages the noisy optimization.

4. Experiments

4.1. Basic Setups

Datasets and Validation Settings. To test the effectiveness of RegionPLC, we evaluate it on three popular datasets:

[1] https://kaldir.vc.in.tum.de/scannet_benchmark/documentation

Method	ScanNet [6]			nuScenes [3]		ScanNet200 [26]	
	B15/N4	B12/N7	B10/N9	B12/N3	B10/N5	B170/N30	B150/N50
3DGenZ [22]	20.6 / 56.0 / 12.6	19.8 / 35.5 / 13.3	12.0 / 63.6 / 6.6	1.6 / 53.3 / 0.8	1.9 / 44.6 / 1.0	2.6 / 15.8 / 1.4	3.3 / 14.1 / 1.9
3DTZSL [4]	10.5 / 36.7 / 6.1	3.8 / 36.6 / 2.0	7.8 / 55.5 / 4.2	1.2 / 21.0 / 0.6	6.4 / 17.1 / 3.9	0.9 / 4.0 / 0.5	0.7 / 3.8 / 0.4
OVSeg-3D [7]	0.0 / 64.4 / 0.0	0.9 / 55.7 / 0.1	1.8 / 68.4 / 0.9	0.6 / 74.4 / 0.3	0.0 / 71.5 / 0.0	1.5 / 21.1 / 0.8	3.0 / 20.6 / 1.6
PLA [7]	65.3 / **68.3** / 62.4	55.3 / 69.5 / 45.9	53.1 / 76.2 / 40.8	47.7 / 73.4 / 35.4	24.3 / 73.1 / 14.5	11.4 / 20.9 / 7.8	10.1 / 20.9 / 6.6
RegionPLC	**69.4** / 68.2 / **70.7**	**68.2 / 69.9 / 66.6**	**64.3 / 76.3 / 55.6**	**64.4 / 75.8 / 56.0**	**49.0 / 75.8 / 36.3**	**16.6 / 21.6 / 13.9**	**14.6 / 22.4 / 10.8**
Fully-Sup.	73.3 / 68.4 / 79.1	70.6 / 70.0 / 71.8	69.9 / 75.8 / 64.9	73.7 / 76.6 / 71.1	74.8 / 76.8 / 72.8	20.9 / 21.7 / 20.1	20.6 / 22.0 / 19.4

Table 2. Results for open-world 3D semantic segmentation on ScanNet, nuScenes and ScanNet200 in terms of hIoU / mIoU$^{\mathcal{B}}$ / mIoU$^{\mathcal{N}}$. Best open-world results are presented in **bold**.

ScanNet [6], ScanNet200 [26] and nuScenes [3], covering indoor and outdoor scenarios. We validate the open-world capability of our method with different numbers of annotated categories, including **base-annotated open world** (*i.e*. part of categories annotated) and **annotation-free open world** (*i.e*. no category annotated). We evaluate our method's performance on both semantic segmentation and instance segmentation tasks.

Category Partition. We split categories into base and novel on ScanNet [6] following PLA [7]. For nuScenes [3], we ignore the "otherflat" class and randomly divide the rest classes into B12/N3 (*i.e*. 12 base and 3 novel categories) and B10/N5. For ScanNet200 [26], we randomly split 200 classes to B170/N30 and B150/N50. See Suppl. for details.

Evaluation Metrics. For semantic segmentation, we follow [7, 34] to employ mIoU$^{\mathcal{B}}$, mIoU$^{\mathcal{N}}$ and harmonic mean IoU (hIoU) for evaluating base, novel categories and their harmonic mean separately. Similarly, for instance segmentation, we employ mAP$^{\mathcal{B}}_{50}$, mAP$^{\mathcal{N}}_{50}$ and hAP$_{50}$.For annotation-free semantic segmentation, we use mean IoU and mean accuracy on foreground classes (*i.e*. mIoU† and mAcc†) excluding "wall", "floor" and "ceiling" for evaluation.

Implementation Details. We adopt the sparse-convolution-based UNet [11] as the 3D encoder with CLIP [25] text encoder as the final classifier for 3D semantic segmentation, and SoftGroup [30] for instance segmentation as [7]. We use category prompts to replace ambiguous category names such as "manmade" and "drivable surface" with a list of concrete category names when encoding category embeddings. We run all experiments with a batch size of 32 on 8 NVIDIA V100 or A100 (see Suppl. for more details).

4.2. Base-annotated Open World

Comparison Methods. We compare RegionPLC to previous open-world or zero-shot works. 3DGenZ [22] and 3DTZSL [4] are early works for 3D zero-shot learning reproduced by [7]. OVSeg-3D extends LSeg to 3D [19], reported by [7]. PLA [7] is the previous cutting-edge method.

3D Semantic Segmentation. As shown in Table 2, compared to the previous state-of-the-art method PLA [7], our method largely lifts the mIoU of unseen categories by $8.3\% \sim 21.8\%$ among various partitions on ScanNet and nuScenes. Furthermore, when compared to baselines without language supervision, *i.e*. 3DGenZ [22] and 3DTZSL [4], our method even obtains $30.6\% \sim 42.7\%$ performance gains regarding mIoU on novel categories among different partitions and datasets. These significant and consistent improvements across indoor and outdoor scenarios show the effectiveness of our RegionPLC framework.

Furthermore, when facing more long-tail dataset ScanNet200 [26], our method still obtains notable mIoU$^{\mathcal{N}}$ gains ranging from 4.2% to 5.1% compared to PLA [7] as shown in Table 2. In this regard, our proposed region-level language supervision and region-aware point-discriminative contrastive loss show its potential to address 3D open-world understanding in complex and long-tail scenarios.

Method	ScanNet		
	B13/N4	B10/N7	B8/N9
OVSeg-3D [19]	5.1 / 57.9 / 2.6	2.0 / 50.7 / 1.0	2.4 / 59.4 / 1.2
PLA [7]	55.5 / 58.5 / 52.9	31.2 / **54.6** / 21.9	35.9 / **63.1** / 25.1
RegionPLC	**58.2 / 59.2 / 57.2**	**40.6** / 53.9 / **32.5**	**46.8** / 62.5 / **37.4**
Fully-Sup.	64.5 / 59.4 / 70.5	62.5 / 57.6 / 62.0	62.0 / 65.1 / 62.0

Table 3. Results for open-world 3D instance segmentation on ScanNet in terms of hAP$_{50}$ / mAP$^{\mathcal{B}}_{50}$ / mAP$^{\mathcal{N}}_{50}$.

3D Instance Segmentation. As our pipeline provides local language descriptions to fine-grained point sets and encourage points to learn discriminative features, it also benefits instance-level localization task. As shown in Table 3, our method consistently brings $4.3\% \sim 12.3\%$ gains compared to the state-of-the-art PLA [7] across three partitions on ScanNet. It is noteworthy that our method obtains more obvious improvements for partitions with fewer base categories (*i.e*. B10/N7 and B8/N9), demonstrating the effectiveness of our RegionPLC in enabling the model to distinguish unseen instances without human annotations.

4.3. Annotation-free Open World

Comparison Methods. As shown in Table 4, we compare two streams of methods: i) Training-free methods using multi-view images for inference [23, 43]. ii) Methods leveraging 2D vision-language models during training [7, 23].

3D Semantic Segmentation. As shown in Table 4, our RegionPLC with SparseUNet32 [11] backbone significantly outperforms all other competitive methods by $1.8\% \sim 57.5\%$ mIoU† and $7.2\% \sim 72\%$ mAcc†. This is the first time that a 3D open-world model achieves state-of-the-art performance without any 3D annotation or 2D pixel-aligned image features but only sparse language supervision for learning. Moreover, our RegionPLC can scale up by scaling the 3D backbone from SparseUNet16 to SparseUNet32, obtaining 2.6% mIoU† gains, which shows the advantage of

Method	Network	mIoU†	mAcc†	Multi-view Infer	GT Instance Mask	Train Hours	Extra Storage	Latency
MaskCLIP‡ [43]	CLIP [25]	23.1	40.9	✓	×	-	-	1.7 s
OpenScene-2D [23]	LSeg [19]	58.0	68.5	✓	×	-	-	106.1s
OpenScene-3D‡ [23]	SparseUNet16 [11]	57.2	69.9	×	×	24.7 h	117.3 G	**0.08 s**
OpenScene-3D‡ [23]	SparseUNet32 [11]	57.8	70.3	×	×	25.3 h	117.3 G	0.10 s
PLA‡ [7]	SparseUNet16 [11]	17.7	33.5	×	×	**11.5 h**	**1.1 G**	**0.08 s**
PLA‡ [7]	SparseUNet32 [11]	19.1	41.5	×	×	12.0 h	**1.1 G**	0.10 s
RegionPLC	SparseUNet16 [11]	56.9	75.6	×	×	12.5 h	5.5 G	**0.08 s**
RegionPLC	SparseUNet32 [11]	**59.6**	**77.5**	×	×	13.0 h	5.5 G	0.10 s
RegionPLC + OpenScene-3D‡	SparseUNet16 [11]	60.1	74.4	×	×	25.9 h	122.8 G	**0.08 s**
RegionPLC + OpenScene-3D‡	SparseUNet32 [11]	**63.6**	**80.3**	×	×	26.4 h	122.8 G	0.10 s
Fully-Sup.	SparseUNet16 [11]	75.9	84.8	×	×	9.6 h	-	0.08 s
Fully-Sup.	SparseUNet32 [11]	**77.9**	**86.2**	×	×	10.5 h	-	0.10 s

Table 4. Annotation-free 3D semantic segmentation on ScanNet. ‡ and $^\sharp$ mean results reproduced by us and Uni3D, independently.

learning from sparse language supervision instead of pixel-aligned feature distillation from 2D encoders [23]. It is also noteworthy that RegionPLC can function as a lightweight plug-and-play module and thus be integrated with other methods such as OpenScene [23] to further boost about 4% mIoU†. Notably, our method is training-efficient, requiring less disk storage and training time compared to OpenScene.

[23]	[7]	RegionPLC	RegionPLC + [23]	Fully-Sup.
5.9 (10.2)	1.8 (3.1)	**9.1** (**17.3**)	**9.6** (**17.8**)	23.9 (32.9)

Table 5. Annotation-free open-world semantic segmentation on ScanNet200 [26] in terms of mIoU† (mAcc†).

Long-tail Scenario. As shown in Table 5, we set up comparisons on the more challenging long-tail dataset ScanNet200 [26]. Notably, our RegionPLC surpasses other counterparts by $3.2\% \sim 7.4\%$ mIoU† and $7.1\% \sim 14.2\%$ mAcc†. Specifically, OpenScene is less effective on ScanNet200 with a large number of fine-grained categories as it inherits the shortcomings or bias of the 2D segmentation model that forgets a large number of concepts during fine-tuning, as verified in [9, 15]. In contrast, our RegionPLC directly learns in a rich vocabulary space with dense and diverse captions which is closer to real open-world scenarios.

4.4. Qualitative Studies

To demonstrate the open-world capability of our RegionPLC, we provide compelling qualitative results showcasing its capability in recognizing and localizing novel categories. As illustrated in Figure 3 (a), RegionPLC successfully identifies numerous categories without any human annotation, demonstrating the quality and richness of our region-level captions and the effectiveness of our region-aware point-discriminative learning objective. For base-annotated cases, our model can recognize challenging tail classes such "keyboard" and "laddar" with precise segmentation in indoor scenarios (see Figure 3 (b)) and small-scale objects with only a few points such as "motorcycle" in outdoor scenarios (see Figure 3 (c)). Moreover, RegionPLC shows a strong localization ability in open-world instance segmentation, accurately grouping novel objects as shown in Figure 3 (d).

5. Ablation Study

In this section, we examine key components of our framework through in-depth ablation studies. Results for base-annotated and annotation-free experiments are measured in hIoU / mIoU$^\mathcal{B}$ / mIoU$^\mathcal{N}$ and mIoU† (mAcc†) separately.

Components					ScanNet	ScanNet
$\mathbf{t}^{v+e}$	$\mathbf{t}^r$	$\mathcal{L}_{\text{pdc}}$	$\mathcal{L}_{\text{rpdc}}$	SFusion	B0/N17	B12/N7
					0.3 (5.3)	24.5 / 70.0 / 14.8
✓					17.7 (33.5)	55.3 / 69.5 / 45.9
	✓				21.7 (37.1)	62.6 / 69.9 / 56.7
	✓	✓			50.6 (71.1)	63.6 / **70.6** / 57.9
	✓		✓		51.7 (72.7)	67.4 / 70.5 / 64.6
	✓		✓	✓	**56.9** (**75.5**)	**68.2** / 69.9 / **66.6**

Table 6. Component analysis on ScanNet. $\mathbf{t}^{v+e}$ and $\mathbf{t}^r$ denotes the combination of view and entity language supervision [7] and best region-level language supervision, respectively.

Component Analysis. Here, we study the effectiveness of our proposed regional captions $\mathbf{t}^r$, the SFusion strategy for caption integration, point-discriminative contrastive loss $\mathcal{L}_{\text{pdc}}$ and its region-aware variants $\mathcal{L}_{\text{rpdc}}$. As shown in Table 6, when compared to view- and entity-level captions used in PLA [7], our region-level language supervision delivers consistent boosts about $4\% \sim 10.8\%$ across different category partitions. Additionally, $\mathcal{L}_{\text{pdc}}$ achieves considerable gains when paired with $\mathbf{t}^r$. Particularly, it brings 28.9% mIoU† gains in the annotation-free setting, illustrating its superiority in learning point-discriminative features for dense parsing tasks. When combined with the region-aware factor, $\mathcal{L}_{\text{rpdc}}$ surpasses $\mathcal{L}_{\text{pdc}}$ by $1.1\% \sim 3.8\%$ mIoU†. Lastly, $2\% \sim 5.2\%$ improvements yielded from SFusion strategy confirm its effectiveness in eliminating redundancy and conflicts from multiple captions for training.

Caption source	$[T_l, T_h]$			
	[0.0, 1.0]*	[0.5, 1.0]	[0.0, 0.5]	[0.0, 0.2]
$\mathbf{t}^{\text{kos}}$, $\mathbf{t}^{\text{sw}}$	53.1 (73.9)	54.6 (75.3)	54.3 (74.4)	56.6 (74.7)
$\mathbf{t}^{\text{kos}}$, $\mathbf{t}^{\text{det-t}}$	54.0 (74.5)	54.1 (73.7)	54.3 (73.9)	55.9 (76.1)
$\mathbf{t}^{\text{kos}}$, $\mathbf{t}^{\text{sw}}$, $\mathbf{t}^{\text{det-t}}$	54.9 (74.2)	55.2 (73.7)	55.4 (75.3)	56.9 (75.5)

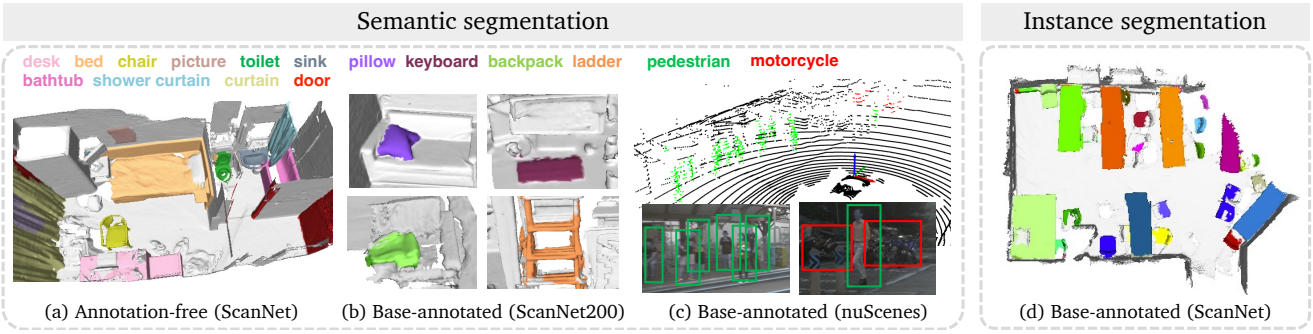
(a) Ablation of various caption sources and caption overlap thresholds. * equals to the data-mixing baseline. We report mIoU (mAcc) here.

$[T_l, T_h]$	[0.5, 1.0]	[0.0, 0.5]	[0.0, 0.5]	[0.0, 0.2]
Ratio (ϵ)	$0.75^\oplus$	$0.34^\oplus$	0.75	$0.72^\oplus$
mIoU (mAcc)	54.6 (75.3)	54.3 (74.4)	55.4 (75.6)	56.6 (74.7)

(b) Ablation of caption overlap thresholds and their ratios when fusing $\mathbf{t}^{\text{kos}}$ and $\mathbf{t}^{\text{sw}}$. $^\oplus$ means the raw ratio for $\mathbf{t}^{\text{kos}}$ and $\mathbf{t}^{\text{kos}}$ + $\mathbf{t}^{\text{sw}}$ with no ϵ applied.

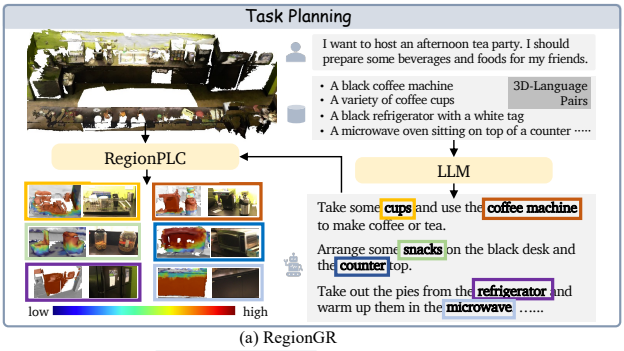
Table 7. SFusion results for zero-shot semantic segmentation considering caption sources, overlap thresholds, and ratios.

SFusion. We also study the effectiveness of our SFusion strategy. As shown in Table 7 (b), with a similar ratio ϵ

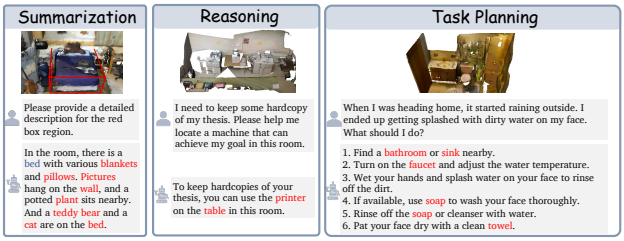


(a) Annotation-free (ScanNet) (b) Base-annotated (ScanNet200) (c) Base-annotated (nuScenes) (d) Base-annotated (ScanNet)

Figure 3. Qualitative results of our RegionPLC. The examples above show annotation-free open-world scene parsing where no human annotation is available (see (a)), and base-annotated open-world learning where a limited number of base classes are annotated (see (b), (c), (d)) for semantic and instance segmentation covering both indoor and outdoor scenarios. Unseen categories are highlighted in colors.



(a) RegionGR



(b) RegionGR with various user queries

Figure 4. (a) Visualizations of RegionGR that integrates LLM for open-ended grounded 3D reasoning. (b) Demonstrating the versatility of RegionGR via more examples of answering user queries.

of the main caption source relative to all merged captions, fusing 3D-language pairs that have a low spatial overlap ratio (i.e., less than 0.5) yields superior results compared to fusing pairs that are highly overlapped (i.e., greater than 0.5). Besides, as shown in Table 7 (a), our SFusion largely outperforms the naive data-mixing strategy ($[T_l, T_h]$=[0.0, 1.0]) with $1.9\% \sim 3.5\%$ gains. These experimental results affirm that potential conflicts and redundancies introduced by regions with a high degree of overlap restrict the benefits derived from multiple language sources. Fortunately, our SFusion method effectively tackles this challenge.

6. Open-ended Grounded 3D Reasoning

Recently, there has been a growing interest in employing language as an interface for connecting human intentions with visual understanding, which facilitates high-level reasoning and planning in the development of embodied agents. Without specific design, RegionPLC can be seamlessly integrated with large language models to enable open-ended grounded 3D reasoning, referred to as RegionGR. As depicted in Figure 4 (a), RegionGR integrates large language models (LLM) for 3D reasoning with regional 3D-language pairs as a knowledge base and utilizes RegionPLC to coarsely locate and identify corresponding objects within the 3D scene for grounded reasoning. Moreover, Figure 4 (b) further exhibits the versatility of RegionGR in responding to user intentions, from summarization to reasoning and planning, particularly within user-specified regions of interest (as shown in the summarization example).

Specifically, RegionGR runs in three steps: (i) We first initialize "environment context" with our *regional 3D-language pairs*, which enables LLM to understand a given scene. If a user specifies an interested 3D region (see Figure 4 (b) left), highly overlapped 3D-language pairs are kept. (ii) We then feed *our prompt in Sec. S4 of suppl.* along with "user query" and "environment context" into LLM to generate answers. (iii) Finally, we parse objects from LLM's response and ground them with RegionPLC.

7. Conclusion

We present RegionPLC, a holistic regional point-language contrastive learning framework to recognize and localize unseen categories in open-world 3D scene understanding. By leveraging advanced VL models and our SFusion strategy, RegionPLC effectively builds comprehensive regional point-language pairs. Furthermore, our region-aware point-discriminative contrastive loss aids in learning distinctive and robust features from regional captions. Extensive experiments demonstrate that RegionPLC remarkably outperforms prior open-world methods in both indoor and outdoor scenarios and excels in challenging long-tail or annotation-free scenarios. Besides, RegionPLC can be effortlessly integrated with LLM for grounded 3D visual reasoning.

Acknowledgement This work has been supported by Hong Kong Research Grant Council - Early Career Scheme (Grant No. 27209621), General Research Fund Scheme (Grant No. 17202422), and RGC Matching Fund Scheme (RMGS). Part of the described research work is conducted in the JC STEM Lab of Robotics for Soft Materials funded by The Hong Kong Jockey Club Charities Trust.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1, 2

[2] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016. 1

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 6

[4] Ali Cheraghian, Shafin Rahman, Dylan Campbell, and Lars Petersson. Transductive zero-shot learning for 3d point cloud classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 923–933, 2020. 6

[5] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 2

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 4, 6

[7] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Language-driven open-vocabulary 3d scene understanding. *arXiv preprint arXiv:2211.16312*, 2022. 1, 2, 4, 6, 7

[8] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Lowis3d: Language-driven open-world instance-level 3d scene understanding. *arXiv preprint arXiv:2308.00353*, 2023. 1, 2

[9] Yuxuan Ding, Lingqiao Liu, Chunna Tian, Jingyuan Yang, and Haoxuan Ding. Don't stop learning: Towards continual learning for the clip model. *arXiv preprint arXiv:2207.09248*, 2022. 7

[10] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017. 2

[11] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 2, 6, 7

[12] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 3

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

[14] Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. *arXiv preprint arXiv:2309.00616*, 2023.

[15] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023. 2, 7

[16] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[17] Maksim Kolodiazhnyi, Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Top-down beats bottom-up in 3d instance segmentation. *arXiv preprint arXiv:2302.02871*, 2023. 2

[18] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8500–8509, 2022. 2

[19] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 1, 2, 6, 7

[20] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1

[21] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *arXiv preprint arXiv:2305.10764*, 2023. 2

[22] Björn Michele, Alexandre Boulch, Gilles Puy, Maxime Bucher, and Renaud Marlet. Generative zero-shot learning for semantic segmentation of 3d point clouds. In *2021 International Conference on 3D Vision (3DV)*, pages 992–1002. IEEE, 2021. 6

[23] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. *arXiv preprint arXiv:2211.15654*, 2022. 1, 2, 4, 6, 7

[24] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 1, 3, 4

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language super-

vision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5, 6, 7

[26] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 125–141. Springer, 2022. 2, 6, 7

[27] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023. 2

[28] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[29] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, Junyeong Kim, and Chang D Yoo. Softgroup++: Scalable 3d instance segmentation with octree pyramid grouping. *arXiv preprint arXiv:2209.08263*, 2022. 2

[30] Thang Vu, Kookhoi Kim, Tung M. Luu, Xuan Thanh Nguyen, and Chang D. Yoo. Softgroup for 3d instance segmentation on 3d point clouds. In *CVPR*, 2022. 2, 6

[31] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052, 2022. 1, 3

[32] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. 1, 3, 4

[33] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 9621–9630, 2019. 2

[34] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019. 6

[35] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3173–3182, 2021. 2

[36] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds. In *Advances in Neural Information Processing Systems*, pages 6737–6746, 2019. 2

[37] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *arXiv preprint arXiv:2209.09407*, 2022. 2

[38] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2019. 2

[39] Andy Zeng, Shuran Song, Stefan Welker, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4238–4245. IEEE, 2018. 1

[40] Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15244–15253, 2023. 1, 2

[41] Junbo Zhang, Runpei Dong, and Kaisheng Ma. Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. *arXiv preprint arXiv:2303.04748*, 2023. 2

[42] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022. 2

[43] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision (ECCV)*, 2022. 6, 7

[44] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 1, 3, 4