PLA: Language-Driven Open-Vocabulary 3D Scene Understanding

Runyu Ding\textsuperscript{1}\textsuperscript{*} Jihan Yang\textsuperscript{1}\textsuperscript{*} Chuhui Xue\textsuperscript{2} Wenqing Zhang\textsuperscript{2} Song Bai\textsuperscript{2}\textsuperscript{‡} Xiaojuan Qi\textsuperscript{1}\textsuperscript{‡}
\textsuperscript{1}The University of Hong Kong \textsuperscript{2}ByteDance

Abstract

Open-vocabulary scene understanding aims to localize and recognize unseen categories beyond the annotated label space. The recent breakthrough of 2D open-vocabulary perception is largely driven by Internet-scale paired image-text data with rich vocabulary concepts. However, this success cannot be directly transferred to 3D scenarios due to the inaccessibility of large-scale 3D-text pairs. To this end, we propose to distill knowledge encoded in pre-trained vision-language (VL) foundation models through captioning multi-view images from 3D, which allows explicitly associating 3D and semantic-rich captions. Further, to foster coarse-to-fine visual-semantic representation learning from captions, we design hierarchical 3D-caption pairs, leveraging geometric constraints between 3D scenes and multi-view images. Finally, by employing contrastive learning, the model learns language-aware embeddings that connect 3D and text for open-vocabulary tasks. Our method not only remarkably outperforms baseline methods by 25.8\% hIoU and 14.5\% \textit{hAP}_{50} in open-vocabulary semantic and instance segmentation, but also shows robust transferability on challenging zero-shot domain transfer tasks. See the project website at https://dingry.github.io/projects/PLA.

1. Introduction

3D scene understanding is a fundamental perception component in real-world applications such as robot manipulation, virtual reality and human-machine interaction. Deep learning has attained remarkable success in this area [13, 38, 28]. However, deep models trained on a human-annotated dataset are only capable of understanding semantic categories in that dataset, \textit{i.e.} close-set prediction. As a result, they fail to recognize unseen categories in the open world (see Fig. 1). This largely restricts their applicability in real-world scenarios with unbounded categories. Besides, heavy annotation costs on 3D datasets (\textit{e.g.} 22.3 minutes for one scene with 20 classes [7]) further make it infeasible to rely on human labor to cover all real-world categories.

This motivates us to study open-vocabulary 3D scene understanding, which equips a model with the ability to localize and recognize open-set classes beyond the label space of an annotated dataset (see Fig. 1). Recently, vision-language (VL) foundation models [33, 22, 47] trained on billions of web-crawled image data with semantic-rich captions [36] are capable of learning adequate vision-language embeddings to connect text and image, which are further leveraged to solve many 2D open-vocabulary tasks including object detection [15, 35], semantic segmentation [43, 26, 51], visual question answering [31] and \textit{etc.} Albeit significantly advancing open-vocabulary image understanding tasks, this pre-training paradigm is not directly viable in the 3D domain due to the absence of large-scale 3D-text pairs.

To this end, initial efforts [50, 20] have attempted to project 3D data into 2D modalities, such as RGB images and depth maps, enabling pre-trained VL foundation models to process the 2D data and achieve object-level open-vocabulary recognition. Nevertheless, this line of methods suffers from several major issues, making it suboptimal to handle scene-level understanding tasks (\textit{e.g.}, instance segmentation). First, multiple RGB images and depth maps are required to represent a 3D sample, which incurs heavy computation and memory costs during training and inference. Second, the projection from 3D to 2D induces information
loss and prohibits direct learning from rich 3D data, leading to subpar performance. Our preliminary study shows the cutting-edge 2D open-vocabulary semantic segmentation method MaskCLIP [51] attains a mere $17.8\%$ mIoU with a 20-fold increase in latency when applied to analyze projected 2D images from 3D ScanNet dataset.

Thus, considering the success of VL foundation models for a variety of vision-language tasks [15, 35, 43, 26, 51, 50, 20], we ask: is it possible to elicit knowledge encoded in powerful VL foundation models to build an explicit association between 3D and language for open-vocabulary understanding? To this end, our core idea is to exploit pre-trained VL foundation models [1, 39] to caption easily-obtained image data aligned with 3D data (i.e. the point set in the corresponding frustum to produce the image). Note that these images can be acquired through neural rendering [9, 46] or from the 3D data collection pipeline [7]. By doing so, we can distill semantic-rich textual descriptions to the 3D domain, which allows explicit association between 3D and vocabulary-rich text for zero-shot 3D scene understanding.

Given 3D-language association, the next question is enabling a 3D network to learn language-aware embeddings from (pseudo) captions. The key challenge stems from intricate object compositions in 3D scene-level data (see Fig. 3), making it difficult to connect objects with corresponding words in the caption. This differs from object-centric image data containing a single centered object [33]. Fortunately, the captioned multi-view images from a 3D scene are related by 3D geometry, which can be leveraged to build hierarchical point-caption pairs, including scene-, view- and entity-level captions. These multi-level point-caption pairs offer coarse-to-fine supervision signals, facilitating learning adequate visual-semantic representations from rich vocabulary by contrastive learning. Without task-specific design, our Point-Language Association paradigm, namely PLA, is generic for various open-vocabulary 3D scene understanding tasks, such as semantic and instance segmentation.

Experimental results for ScanNet [7] and S3IDS [2] datasets show the effectiveness of our method in in-domain open-vocabulary tasks with only category shifts, i.e. training and evaluation are conducted on the same dataset, surpassing baselines by $25.8\% \sim 44.7\%$ hIoU on semantic segmentation and $14.5\% \sim 50.4\%$ hAP$_{50}$ on instance segmentation. Besides, our model, trained on a dataset (i.e. ScanNet), can generalize to another dataset (i.e. S3IDS) with both data distribution and category shifts, manifesting its transferability. Finally, our model can benefit from more advanced foundation models that provide higher-quality caption supervision, showing its scalability and extensibility.

2. Related Work

3D scene understanding focuses on understanding the semantic meaning of objects and surrounding environment from point clouds. In this work, we focus on two fundamental scene understanding tasks: semantic and instance segmentation. 3D semantic segmentation aims to obtain point-wise semantic predictions for point clouds. Representative works develop point-based solutions [32, 19] with elaborately designed point convolution operations [37, 42] or transformers [24] or voxel-based [13, 6] methods with 3D sparse convolutions [14] to produce point-wise segmentation results. 3D instance segmentation further targets distinguishing different object instances based on semantic segmentation. Existing approaches either adopt a top-down solution [45, 44] via predicting 3D bounding box followed by mask refinement, or a bottom-up [23, 38] approach through grouping points. However, existing methods cannot recognize open-set novel categories, which we aim to address.

Zero-shot and open-vocabulary understanding aims to recognize novel classes that are not annotated in training data. Early approaches mainly follow zero-shot settings that can be coarsely grouped into discriminative methods [40, 3] and generative methods [4, 16]. 3DGenZ [27] extends [4] to the 3D scenario for zero-shot semantic segmentation. Going beyond zero-shot learning, the more general open-vocabulary setting assumes a large vocabulary corpus is accessible during training [49]. Existing 2D open-vocabulary learning works either exploit massive annotated image-text pairs to provide weak supervision for expanding vocabulary size [49, 53] or leverage pre-trained VL models from large-scale image-caption pairs, such as CLIP [33], to address open-vocabulary recognition where knowledge distillation [35, 15, 48] and prompt learning [12, 11] are studied.

In comparison, 3D open-vocabulary learning is still in its infancy with only a few explorations focusing on object classification [50, 20]. They attempt to project object-level 3D point clouds to multi-view 2D images and depth maps to adopt the pre-trained VL model to generate open-vocabulary predictions, which, however, suffer from heavy computation and poor performance if applied to 3D scene understanding tasks. In this work, we propose a language-driven 3D open-vocabulary framework that directly associates 3D with text descriptions leveraging multi-view images and VL foundation models. It can be generally applied to various scene understanding tasks and is efficient with only the 3D network employed in training and inference.

3. Method

3.1. Preliminary

3D open-vocabulary scene understanding aims to localize and recognize unseen categories without corresponding human annotation as supervision. Formally, annotations on semantic and instance levels $Y = \{y^\text{sem}, y^\text{ins}\}$ are divided into base $C^B$ and novel $C^N$ categories. In the training stage, the 3D model can access all point clouds $P = \{p\}$ but only annotations for base classes $Y^B$, unaware of both annotations $Y^N$ and category names concerning novel classes $C^N$.\n
7011
Figure 2. Our language-driven 3D scene understanding paradigm. Different from the close-set network, the learnable semantic head is replaced by category embeddings encoded by a text encoder from category names. Binary head is to rectify semantic scores with base and novel probability as conditions. Instance head is tailored to instance segmentation. Most importantly, to endow the model with rich semantic space to improve open-vocabulary capability, we supervise point embeddings with caption embeddings based on point-language association (see Fig. 3 for details). Best viewed in color.

However, during inference, the 3D model needs to localize objects and classify points belonging to both base and novel categories.

As for a typical scene understanding network, it consists of a 3D encoder $F_{3D}$, a dense semantic classification head $F_{sem}$, and an instance localization head $F_{loc}$ (see Suppl. for details). Its inference pipeline can be demonstrated below:

$$f^p = F_{3D}(p), \quad s = \sigma \circ F_{sem}(f^p), \quad z = F_{loc}(f^p, s),$$

where $p$ is the input point cloud, $f^p$ is point-wise visual feature, $s$ is semantic score, $z$ is the instance proposal output and $\sigma$ is the softmax function. With these network predictions, we can then calculate semantic classification loss $L_{sem}$ with semantic label $y^sem$, and localization loss $L_{loc}$ with instance label $y^{ins}$ similar to [23, 38] as Eq (2).

$$L_{sem} = \text{Loss}(s, y^{sem}), \quad L_{loc} = \text{Loss}(z, y^{ins}).$$

### 3.2. Open-Vocabulary Setups

Though we can train a scene understanding model with loss functions in Eq. (2), it is actually a close-set model with a close-set classifier $F_{sem}$, incapable of recognizing unseen categories. In this regard, we introduce the text-embedded semantic classifier to obtain an open-vocabulary model and propose a binary calibration module to correct the bias toward base categories for open-vocabulary inference.

#### 3.2.1 Text-Embedded Semantic Classifier

First, as shown in Fig. 2, to make the model become an open-vocabulary learner, we replace its learnable semantic classifier $F_{sem}$ with category embeddings $f^l$ and a learnable vision-language adapter $F_{\theta}$ to match the dimension between 3D features $f^p$ and $f^l$ as follows,

$$f^v = F_{\theta}(f^p), \quad s = \sigma(f^l \cdot f^v),$$

where $f^v$ is the projected features with the VL adapter $F_{\theta}$, $f^l = [f^l_1, f^l_2, \ldots, f^l_C]$ is a series of category embeddings obtained by encoding category names $C$ with a frozen text encoder $F_{\text{text}}$ such as BERT [10] or CLIP [33] (see Fig. 2).

The prediction is made by calculating the cosine similarity among projected point features $f^v$ and categories $f^l$ and then selecting the most similar category. Notice that $f^l$ only contains embeddings belonging to base categories $C^B$ during training, but embeddings related to both base and novel classes $C^B \cup C^N$ are used during open-vocabulary inference.

With category embeddings $f^l$ as a classifier, the model can support open-vocabulary inference with any desired categories. The above design generally follows LSeg [26] and is named LSeg-3D as a baseline.

#### 3.2.2 Semantic Calibration with Binary Head

Although the model has already possessed open-vocabulary capability, we empirically find that it can hardly make any correct predictions on novel classes but mistakes them for base classes. As the model is only trained to recognize base categories, it inevitably produces over-confident predictions on base classes regardless of their correctness, also known as the calibration problem [17]. To this end, we propose a binary calibration module to rectify semantic scores with the probability of a point belonging to base or novel classes.

Specifically, as shown in Fig. 2, the binary head $F_b$ is employed to distinguish annotated (i.e., base) and unannotated (i.e., novel) points. During training, $F_b$ is optimized with:

$$s^b = F_b(f^p), \quad L_{bi} = \text{BCELoss}(s^b, y^b),$$

where $\text{BCELoss}(\cdot, \cdot)$ is the binary cross-entropy loss, $y^b$ is the binary label and $s^b$ is the predicted binary score indicating the probability that a point belongs to novel categories.

In the inference stage, we then exploit the binary probability $s^b$ to correct the over-confident semantic score $s$ as follows,

$$s = s_B \cdot (1 - s^b) + s_N \cdot s^b,$$

where $s_B$ is the semantic score computed solely on base classes with novel class scores set to zero. Similarly, $s_N$ is computed only for novel classes, setting base class scores to zero. We empirically show that the probability calibration largely improves the performance of both base and novel categories (see Sec. 5), demonstrating that our design effectively corrects over-confident semantic predictions.

### 3.3. Image-Bridged Point-Language Association

With the text-embedded classifier and the binary semantic calibration module, we obtain a deep model with...
open-vocabulary capability. Nevertheless, its performance on novel categories is very close to random guesses as shown in Table 5. Recent success of open-vocabulary works [26, 35, 15] in 2D vision community shows the effectiveness of introducing language supervision to guide vision backbones. Language supervision can not only enable the vision backbone to access abundant semantic concepts with a large vocabulary size but also assist in mapping vision and language features into a common space to facilitate multi-modality downstream tasks. However, Internet-scale paired point-text data are not as readily available as image-text pairs on social media, which largely hinders the development of language-driven 3D understanding.

To address this challenge, we propose PLA, an image-bridged point-language association module to provide language supervision for 3D scene perception without human annotation (see Fig. 2 & Fig. 3). Our core idea is to use multi-view images of a 3D scene as a bridge to access knowledge encoded in VL foundation models. As shown in Fig. 3, a text description is first generated by a powerful image-captioning model taking images of 3D scenes as input, and then associated with a set of points in the 3D scene using the projection matrix between images and 3D scenes. We elaborate on our captioning procedure as well as the designed hierarchical point-caption association as follows.

### 3.3.1 Caption Multi-View Images

As image captioning is a fundamental task in VL research area [18], various foundation models [39, 1, 29] trained with massive samples are readily available for solving this task. Specifically, taking the $j^{th}$ image of the $i^{th}$ scene $v_{ij}$ as input, the pre-trained image-captioning model $G$ can generate its corresponding language description $t^v_{ij}$ as follows,

$$t^v_{ij} = G(v_{ij}).$$

Surprisingly, though $G$ has not been specifically trained on the 3D scene understanding dataset, the entities in generated captions already cover the whole semantic label space of the popular 3D scene understanding dataset ScanNet [7]. In addition, the caption $t$ provides fairly accurate and comprehensive descriptions for room types, semantic categories with color and texture attributes, and even spatial relations (see language supervision $\{t^v\}$ examples in Fig. 3 and more examples in Suppl.).

### 3.3.2 Associate Point Cloud with Language

Given the image-caption pairs, the next step is to connect a point set $p$ to language $t$ with images $v$ as bridge as follows:

$$\text{Explore } \langle \hat{p}, v \rangle \text{ with } \langle p, v \rangle \text{ and } \langle v, t \rangle.$$  

Here, we propose three association fashions on point sets with different spatial scales.

#### Scene-Level Point-Caption Association

The simplest and coarsest association manner is to link language supervision to all points in a given 3D point cloud scene $\hat{p}^v = p$. As illustrated in Fig. 3, we take all 2D image captions $t_{ij}^v$ of a given scene $p_j$ to obtain a scene-level caption $t_j^v$ via a text summarizer $G\text{sum}$ as follows:

$$t_j^v = G\text{sum}(\{t_{1j}^v, t_{2j}^v, \ldots, t_{n_j}^v\}),$$

where $n_j$ is the number of images of scene $p_j$. By forcing each scene $p$ to learn from the corresponding scene descriptions $t^v$, abundant vocabulary and visual-semantic relationships are introduced to improve the language understanding capability of a 3D network. Despite the simplicity of scene-level caption, we empirically find that it can lift the model’s open-vocabulary capability by a large margin (see Sec. 5).

#### View-Level Point-Caption Association

Albeit effective, scene-level caption only provides a single caption for all points in a scene, which overlooks the relation of language to local 3D point clouds, rendering it sub-optimal for scene understanding tasks. In this regard, we further propose a view-level point-caption association that leverages the geometrical relationship between image and points to assign each image caption $t^v$ with a point set inside the 3D view frustum $\hat{p}^v$ of the given image $v$ (see blue box in Fig. 3). Specifically, to obtain the view-level point set $\hat{p}^v$, we first back-project the RGB image $v$ to 3D space using the depth information $d$ to get its corresponding point set $\hat{p}$:

$$[ \hat{p} | \hat{v} | \hat{d} ] = T^{-1} [ v | d ],$$

where $[\cdot|\cdot]$ denotes block matrix, $T \in \mathbb{R}^{3 \times 4}$ is the projection matrix comprising of camera intrinsic matrix and rigid transformations obtained by sensor configurations or mature SLAM approaches [8]. As back-projected points $\hat{p}$ and
points in 3D scene \( \mathbf{p} \) may be only partially overlapped, we then compute their overlapped regions to get the view-level point set \( \hat{\mathbf{p}}^v \) as follows,

\[
\hat{\mathbf{p}}^v = V^{-1}(R(V(\hat{\mathbf{p}}), V(\mathbf{p}))),
\]

where \( V \) and \( V^{-1} \) are the voxelization and reverse-voxelization processes, and \( R \) denotes the radius-based nearest-neighbor search [52]. Such a view-based association enables the model to learn with region-level language description, which largely strengthens the model’s recognition and localization ability on unseen categories.

**Entity-Level Point-Caption Association.** Although view-level caption can already associate each image-caption \( t^v \) with a concrete partial point set in a 3D scene, such an association still constructs on a large 3D area (i.e. around 25K points) with multiple semantic objects/categories as shown in Fig. 3. This is not friendly for the 3D network to learn fine-grained point-wise semantic attributes and instance-wise position information from caption supervision. In this regard, we further propose a fine-grained point-language association that owns the potential to build entity-level point-caption pairs, i.e. object instances with a caption.

Specifically, as illustrated in Fig. 3, we leverage the differences and intersections of adjacent view-level point sets \( \hat{\mathbf{p}}^v \) and their corresponding view-caption \( t^v \) to obtain the entity-level associated points \( \hat{\mathbf{p}}^e \) and caption \( t^e \). First, we calculate entity-level caption \( t^e \) as below:

\[
w_{ij} = E(t^e_i),
\]

\[
w_{ij} = w_i \setminus w_j, \quad w_{ij}^e = w_i \setminus w_j, \quad w_{ij}^v = w_i \cap w_j, \quad t^v = \text{Concat}(w^v),
\]

where \( E \) denotes extracting a set of entity words \( w \) from caption \( t^v \), and \( \setminus \) and \( \cap \) represent the set difference and intersection separately, and Concat denotes the concatenation of all words with spaces to form an entity-level caption \( t^e \).

Similarly, we can easily calculate entity-level point sets and associate them to previously obtained entity-level captions to form point-caption pairs as below:

\[
\hat{\mathbf{p}}^e_{ij} = (\hat{\mathbf{p}}^e_i \setminus \hat{\mathbf{p}}^e_j), \quad \hat{\mathbf{p}}^v_{ij} = (\hat{\mathbf{p}}^v_i \setminus \hat{\mathbf{p}}^v_j), \quad \hat{\mathbf{p}}^e_{ij} = \hat{\mathbf{p}}^e_i \cap \hat{\mathbf{p}}^e_j, \quad \hat{\mathbf{p}}^v_{ij} = \hat{\mathbf{p}}^v_i \cap \hat{\mathbf{p}}^v_j,
\]

\[
< \hat{\mathbf{p}}^e_{ij}, t^v_{ij} >, < \hat{\mathbf{p}}^v_{ij}, t^e_{ij} >, < \hat{\mathbf{p}}^v_{ij}, t^v_{ij} >.
\]

With entity-level \((\hat{\mathbf{p}}^e, t^e)\) pairs, we further filter them to ensure each entity-level point set \( \hat{\mathbf{p}}^e \) relates to at least one entity and focuses on a small enough 3D space as follows,

\[
\gamma < |\hat{\mathbf{p}}^e| < \delta \cdot \min(|\hat{\mathbf{p}}^e|, |\hat{\mathbf{p}}^v|) \quad \text{and} \quad |t^e| > 0,
\]

where \( \gamma \) is a scalar to define minimal number of points, \( \delta \) is a ratio to control the maximum size of \( \hat{\mathbf{p}}^e \), and caption \( t^e \) is not empty. Such a constraint helps focus on a fine-grained 3D space with fewer entities in each caption supervision.

**Comparison Among Different Point-Caption Association Manners.** The above-proposed three coarse-to-fine point-caption association manners actually hold different merits and drawbacks. As shown in Table 1, the scene-level association has the simplest implementation but obtains the coarsest correspondence between captions and points (i.e. each caption corresponds to over 140K points); the view-level association provides point-language mapping relation at a finer level, enjoying a larger semantic label space (i.e. over 20× more captions) and a more localized point set (i.e. around 6× fewer corresponding points per caption) than scene caption; the entity-level association owns the most fine-grained correspondence relation, matching each caption to only 4K points on average, and thus can further benefit dense prediction and instance localization in downstream tasks. We empirically show that the fine-grained association and the semantic-rich label space are two important factors for open-vocabulary perception tasks (see Sec. 5).

### 3.4. Contrastive Point-Language Training

With obtained point-caption pairs \((\hat{\mathbf{p}}, t)\), we are ready to guide the 3D network \( F_{3D} \) to learn from vocabulary-rich language supervision. Here, we introduce a general point-language feature contrastive learning that can be applied to all kinds of coarse-to-fine point-caption pairs.

Specifically, we first obtain caption embeddings \( f^t \) with a pre-trained text encoder \( F_{\text{text}} \). As for the associated partial point set \( \hat{\mathbf{p}} \), we select its corresponding point-wise features from adapted features \( f^v \) and leverage global average pooling to obtain its feature vector \( f^p \) as follows,

\[
f^t = F_{\text{text}}(t), \quad f^p = \text{Pool}(\hat{\mathbf{p}}, f^v).
\]

We then adopt contrastive loss as [49] to pull corresponding point-caption feature embeddings closer and push away unrelated point-caption features as follows,

\[
L_{\text{cap}} = -\frac{1}{n} \sum_{i=1}^{n} \log \frac{\exp(f^t_i \cdot f^v_i / \tau)}{\sum_{j=1}^{n} \exp(f^t_i \cdot f^v_j / \tau)},
\]

where \( n \) is the number of point-caption pairs in any given association fashion and \( \tau \) is a learnable temperature to modulate the logits as CLIP [33]. It is also worth noting that we remove duplicate captions in a batch to avoid noisy optimization during contrastive learning. With Eq. (17) and Eq. (18), we can easily compute caption losses on scene-level \( L_{\text{cap}} \), view-level \( L_{\text{cap}} \) and entity-level \( L_{\text{cap}} \). Our final caption loss is a weighted combination as follows,

\[
L_{\text{cap}} = \alpha_1 L_{\text{cap}} + \alpha_2 L_{\text{cap}} + \alpha_3 L_{\text{cap}},
\]

where \( \alpha_1, \alpha_2 \) and \( \alpha_3 \) are trade-off factors. As shown in Fig. 2, the overall training objective can be written as

\[
\mathcal{L} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{loc}} + L_{\text{all}} + \mathcal{L}_{\text{bi}}.
\]
Table 2. Results for open-vocabulary 3D semantic segmentation on ScanNet and S3DIS in terms of hIoU, mIoU and hIoU mIoU.

<table>
<thead>
<tr>
<th>Method</th>
<th>C^N prior</th>
<th>ScanNet</th>
<th>S3DIS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>B15/N4 hIoU mIoU hIoU mIoU</td>
<td>B10/N9 hIoU mIoU hIoU mIoU</td>
</tr>
<tr>
<td>LSeg-3D [26]</td>
<td>×</td>
<td>00.0 64.4 00.0</td>
<td>00.0 49.0 00.0</td>
</tr>
<tr>
<td>3DGenZ [27]</td>
<td>✓</td>
<td>20.6 56.0 12.6</td>
<td>12.0 63.6 06.6</td>
</tr>
<tr>
<td>3DTZSL [5]</td>
<td>✓</td>
<td>01.8 68.4 00.9</td>
<td>00.8 43.1 04.7</td>
</tr>
<tr>
<td>PLA (w/o Cap.)</td>
<td>×</td>
<td>39.7 63.8 28.0</td>
<td>25.7 75.6 15.5</td>
</tr>
<tr>
<td>PLA</td>
<td>×</td>
<td>65.3 63.8 62.4</td>
<td>53.1 76.2 40.8</td>
</tr>
<tr>
<td>PLA (w/ self-train)</td>
<td>✓</td>
<td>70.3 68.9 71.7</td>
<td>59.2 76.9 48.2</td>
</tr>
<tr>
<td>Fully-Sup.</td>
<td>✓</td>
<td>73.3 68.4 79.1</td>
<td>69.9 75.8 64.9</td>
</tr>
</tbody>
</table>

Table 3. Results for open-vocabulary 3D instance segmentation on ScanNet and S3DIS in terms of hAP, mAP and mAP_50.

<table>
<thead>
<tr>
<th>Method</th>
<th>C^N prior</th>
<th>ScanNet</th>
<th>S3DIS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>B15/N4 hAP mAP_50 mAP_50</td>
<td>B10/N7 hAP mAP_50 mAP_50</td>
</tr>
<tr>
<td>LSeg-3D [26]</td>
<td>×</td>
<td>05.1 57.9 02.6</td>
<td>02.0 50.7 01.0</td>
</tr>
<tr>
<td>PLA (w/o Cap.)</td>
<td>×</td>
<td>21.0 59.6 12.6</td>
<td>15.9 63.2 09.1</td>
</tr>
<tr>
<td>PLA</td>
<td>×</td>
<td>55.5 58.5 52.9</td>
<td>31.2 54.6 21.9</td>
</tr>
<tr>
<td>PLA (w/ self-train)</td>
<td>✓</td>
<td>58.6 58.0 59.2</td>
<td>41.4 56.9 32.6</td>
</tr>
<tr>
<td>Fully-Sup.</td>
<td>✓</td>
<td>64.5 59.4 70.5</td>
<td>62.5 57.6 62.0</td>
</tr>
</tbody>
</table>

4. Experiments

4.1. Basic Setups

Datasets and Perception Tasks. To validate the effectiveness of our point-language association paradigm, we conduct experiments on two datasets: ScanNet [7] densely annotated in 20 classes and S3DIS [2] with 13 classes on both semantic and instance segmentation tasks.

Category Partitions. Without standard open-vocabulary partitions on these two datasets, we build an open-vocabulary benchmark with multiple base/novel partitions. To circumvent model confusion, we disregard the “otherfurniture” class in ScanNet and the “clutter” class in S3DIS as they lack exact semantic meanings and can include any semantic categories. As for ScanNet, we randomly partition the rest 19 classes into 3 base/novel partitions for semantic segmentation, i.e., B15/N4, B12/N7 and B10/N9, where B15/N4 indicates 15 base and 4 novel categories. We also follow SoftGroup [38] to exclude two background classes and thus obtain B13/N4, B10/N7, and B8/N9 partitions for instance segmentation on ScanNet. As for S3DIS, we randomly shuffle the rest 12 classes into 2 base/novel splits, i.e., B8/N4, B6/N6 for both semantic and instance segmentation. Specific category splits are presented in the Suppl.

Metrics. We employ widely adopted mean intersection over union (mIoU) and mean average precision under 50% IoU threshold (mAP_50) as evaluation metrics for semantic and instance segmentation, respectively. These metrics are calculated on base and novel classes separately with superscripts of B and N (e.g., mIoU_50). Further, we use harmonic mean IoU (hIoU) and AP_50 (hAP_50) as major indicators following popular zero-shot learning works [40, 43] to consider category partition between base and novel.

Architectures and Baseline Methods. We adopt the popular and high-performance sparse convolutional UNet [13, 6] as 3D encoder F_D, the text encoder of CLIP as F_text, two fully-connected layers with batch normalization [21] and ReLU [30] as VL adapter F_v, an UNet decoder as binary head F_ω. Also, we utilize the state-of-the-art instance segmentation network SoftGroup [38] for instance head F_ins.

As for baseline methods, other than the above-mentioned LSeg-3D in Sec.3.2.1, we also re-produce two 3D zero-shot learning methods 3DGenZ [27] and 3DTZSL [5] with task-tailored modifications. The implementation details are provided in the Suppl.

4.2. Main Results

3D Semantic Segmentation. As shown in Table 2, compared to LSeg-3D [26] baseline, our method obtains around 51.3% ~ 65.3% and 34.5% ~ 38.5% hIoU improvements among different partitions on ScanNet and S3DIS respectively, demonstrating its superior open-vocabulary capability. Even compared to previous zero-shot methods 3DGenZ [27] and 3DTZSL [5] that know novel category names during training, our method still obtains 35.5% ~ 54.8% improvements in terms of hIoU among various partitions on ScanNet. Especially, our PLA trained model largely surpasses its no caption supervision counterparts (i.e., PLA (w/o Cap.)) by 25.6% ~ 30.8% hIoU and 21.6% ~ 26.3% hIoU on ScanNet and S3DIS, respectively. It is noteworthy that the improvement from our method is consistent on different base/novel partitions and datasets, further illustrating its robustness and effectiveness.

3D Instance Segmentation. As demonstrated in Table 3, our method remarkably surpasses baseline methods by 29.2% ~ 50.4% hAP_50 and 14.5% ~ 14.9% hAP_50 among different base/novel partitions on ScanNet and S3DIS, re-
respectively. Such outstanding performance indicates our contrastive point-language training helps the 3D backbone learn not only semantic attributes but also instance localization information from captions. Notice that the improvement for S3DIS is slightly larger than ScanNet on both semantic segmentation and instance segmentation. This is actually caused by S3DIS’s smaller number of training samples (only 271 scenes) and much fewer point-caption pairs owing to fewer overlapped regions between images and 3D scenes.

Self-Bootstrap with Novel Category Prior. As some existing zero-shot methods (i.e., 3DGenZ [27] and 3DTZSL [5]) can access novel category names but no human-annotation during training, here we also provide a simple variant to leverage such novel category prior in self-training fashion [41]. As shown in Table 2 and 3, PLA (w/ self-train) obtains around 2% ~ 12% gains among semantic and instance segmentation on two datasets. This demonstrates that our model can further self-bootstrap its zero-shot capability and extend its vocabulary size without any human annotation.

### 4.3. Zero-shot Domain Transfer

Our method already shows excellent potential in solving in-domain open-vocabulary scene understanding tasks with category shifts. However, transferable open-vocabulary learners across different domains/datasets also merit exploration, as they face both category and data distribution shifts. In this regard, we conduct zero-shot domain transfer experiments that train the model on ScanNet’s base classes and test it on all S3DIS classes without fine-tuning. Notably, S3DIS has 4 categories not present in ScanNet. As shown in Table 4, our PLA consistently outperforms LSeg-3D [26] by 7.7% ~ 18.3% mIoU for semantic segmentation and 5.0% ~ 9.5% mAP50 for instance segmentation. Such outstanding improvements substantiate our model’s generality for both category shift and data distribution shift. Note that we do not use the binary head for domain transfer here, as the base/novel partition is dataset-specific. We leave calibrating base and novel semantic predictions in out-of-domain open-vocabulary scenarios to future work.

### 5. Ablation Studies

In this section, we examine key components of our framework through in-depth ablation studies. Experiments are conducted on ScanNet B15/N4 partition by default. The default setting is marked in gray.

<table>
<thead>
<tr>
<th>Components</th>
<th>hIoU / mIoU / mIoU(^{50}) / mAP50</th>
<th>hAP50 / mAP50 / mAP50 (^{50})</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>00.0 / 64.4 / 00.0</td>
<td>05.1 / 57.9 / 02.6</td>
</tr>
<tr>
<td>✓</td>
<td>39.8 / 68.5 / 28.1</td>
<td>41.0 / 59.6 / 12.8</td>
</tr>
<tr>
<td>✓</td>
<td>54.6 / 67.8 / 45.1</td>
<td>52.8 / 57.8 / 36.6</td>
</tr>
<tr>
<td>✓</td>
<td>61.3 / 68.5 / 55.5</td>
<td>55.9 / 58.9 / 53.3</td>
</tr>
<tr>
<td>✓</td>
<td>63.6 / 67.8 / 60.0</td>
<td>56.6 / 59.0 / 54.4</td>
</tr>
<tr>
<td>✓</td>
<td>61.9 / 68.1 / 56.8</td>
<td>54.9 / 59.5 / 51.0</td>
</tr>
<tr>
<td>✓</td>
<td>65.3 / 68.3 / 62.4</td>
<td>55.5 / 58.5 / 52.9</td>
</tr>
<tr>
<td>✓</td>
<td>64.6 / 69.0 / 60.8</td>
<td>54.5 / 58.2 / 51.4</td>
</tr>
</tbody>
</table>

Table 5. Component analysis on ScanNet. Binary denotes binary head calibration. Cap\(^a\), Cap\(^b\) and Cap\(^c\) denotes scene-level, view-level and entity-level caption supervision, respectively.

**Caption Composition Analysis.** As a caption can composite entities (e.g., sofa), their relationships (e.g., spatial relation) and attributes (e.g., color and texture), we investigate which types of words mainly contribute to the open-vocabulary capability. As shown in Table 6, when only keeping entity phrases in the caption, (a) variant even outperforms the full caption variant. In addition, if we only keep entities that exactly match category names in captions, obtained (b) variant suffers over 13% mIoU degradation on novel categories, showing that diverse entity words to expand semantic space is a crucial factor for captions. Furthermore, although the (c) variant introduces both correct base and novel label names in the caption, it still obtains slightly inferior performance to our foundation-model-generated caption, illustrating existing foundation models are powerful enough to provide promising supervision.

<table>
<thead>
<tr>
<th>Caption Composition</th>
<th>hIoU / mIoU(^{50}) / mIoU(^{50})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) keep only entities</td>
<td>65.7 / 69.0 / 62.7</td>
</tr>
<tr>
<td>(b) keep only label names</td>
<td>57.6 / 68.5 / 49.6</td>
</tr>
<tr>
<td>(c) ground-truth label names</td>
<td>64.8 / 68.1 / 61.9</td>
</tr>
<tr>
<td>(d) full caption</td>
<td>65.3 / 68.3 / 62.4</td>
</tr>
</tbody>
</table>

Table 6. Ablation of caption composition.

**Text Encoder Selection.** Here, we compare different text
encoders \( F_{\text{text}} \) for extracting caption and category embeddings. As shown in Table 7, the vision-language pre-trained text encoder of CLIP [33] shows over 7% higher mIoU than BERT [10] and GPT2 [34] that are only pre-trained on language modality. This demonstrates that the vision-aware text encoder can provide better language embedding for 3D-language tasks since 3D also leverages texture, shape and RGB information as images for recognition.

Foundation Model for Image Captioning. By default, we employ one of the most popular open-source image captioning models, GPT-ViT2 [1], on the HuggingFace platform to generate captions in main experiments. However, as shown in Table 8, the recent state-of-the-art foundation model OFA [39] can consistently surpass GPT-ViT2 on three partitions, which reflects the potential of our method to be further boosted with stronger foundation models.

### 6. Qualitative Analysis

To more straightforwardly illustrate the open-vocabulary ability of our method, we present some interesting qualitative results in terms of recognizing synonymical classes, abstract classes and even unannotated classes.

**Synonymical Novel Classes.** Here, we substitute class names with related but new words for inference. As illustrated in Fig. 4 (a), when we replace “sofa” with “couch” or “refrigerator” with “freezer”, the model still attains a high-quality segmentation mask. This demonstrates our model is robust to recognize synonymical concepts.

**Abstract Novel Classes.** Apart from object entities, we find the model is able to understand more abstract concepts such as room types. As shown in Fig. 4 (b), by removing “shower curtain”, “toilet”, “sink” and “bathtub” in input categories and adding “bathroom”, the predicted “bathroom” roughly covers the real bathroom region. The right example shows the model can also understand ‘kitchen’ regions. It indicates our model is capable to recognize out-of-vocabulary and abstract concepts beyond concrete semantic objects.

**Unannotated Novel Classes.** As current 3D datasets fail to annotate all classes due to insufferable annotation costs, our model owns the potential to recognize those unannotated classes with high-quality predictions, facilitating open-world applications. As shown in Fig. 4 (c), the model successfully identifies “monitor” and “blackboard” that are not included in the dataset annotations with accurate masks.

### 7. Conclusion

We propose PLA, a general and effective language-driven 3D scene understanding framework that enables the 3D model to localize and recognize novel categories. By leveraging images as a bridge, we construct hierarchical point-language pairs harvesting powerful 2D VL foundation models and geometric constraints between 3D scenes and 2D images. We employ contrastive learning to pull features of such associated pairs closer, introducing rich semantic concepts into the 3D network. Extensive experimental results show the superiority of our method on not only in-domain open-vocabulary semantic and instance segmentation, but also challenging out-of-domain zero-shot transfer.

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 JCR STEMLab of Robotics for Soft Materials funded by The Hong Kong Jockey Club Charities Trust.
References


[26] Boyi Li, Kilian Q Weinberger, Sergei Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic seg-


