LASS3D: Language-Assisted Semi-Supervised 3D Semantic Segmentation with Progressive Unreliable Data Exploitation

Jianan Li[1,2] and Qiulei Dong *[1,2,3]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[3] Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences
lijianan211@mails.ucas.ac.cn, qldong@nlpr.ac.cn

Abstract. Precisely annotating large-scale 3D datasets for point cloud segmentation is laborious. To alleviate the annotation burden, several semi-supervised 3D segmentation methods have been proposed in literature. However, two issues remain to be tackled: 1) The utilization of large language-vision models (LVM) in semi-supervised 3D semantic segmentation remains under-explored. 2) The unlabeled points with low-confidence predictions are directly discarded by existing methods. Taking these two issues into consideration, we propose a language-assisted semi-supervised 3D semantic segmentation method named **LASS3D**, which is built upon the commonly used MeanTeacher framework. In LASS3D, we use two off-the-shelf LVM to generate multi-level captions and leverage the images as the bridge to connect the text data and point clouds. Then, a semantic-aware adaptive fusion module is explored in the student branch, where the semantic information encoded in the embeddings of multi-level captions is injected into 3D features by adaptive fusion and then the semantic information in the text-enhanced 3D features is transferred to the teacher branch by knowledge distillation. In addition, a progressive exploitation strategy is explored for the unreliable points in the teacher branch, which can effectively exploit the information encapsulated in unreliable points via negative learning. Experimental results on both outdoor and indoor datasets demonstrate that LASS3D outperforms the comparative methods in most cases.

Keywords: 3D semantic segmentation · Semi-supervised segmentation

1 Introduction

3D Semantic Segmentation [7,22–24,42,43,51] is an important task for perceiving real-world environments, and intensive efforts have been devoted to this task in recent years. However, most of the existing works [8,22,41,42,51,54] in literature

* Corresponding author

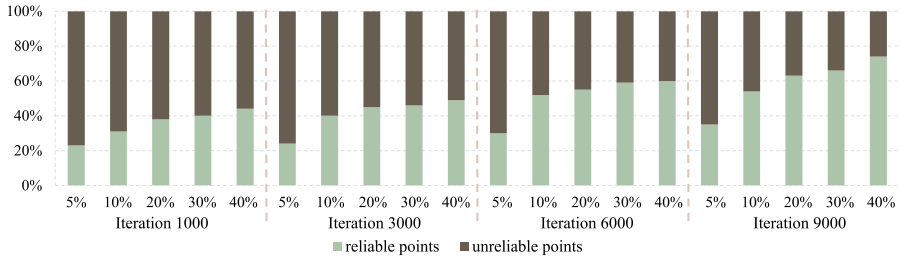


Fig. 1: Illustration of the proportions of reliable and unreliable points predicted by GPC [17] on ScanNet [4]. The horizontal axis denotes the labeled ratio and training iteration. The vertical axis denotes the proportion.

are fully-supervised methods, which require time-consuming and labor-intensive data annotation for training.

This issue encourages researchers to investigate semi-supervised 3D semantic segmentation, where only a small amount of densely labeled data and a large amount of unlabeled data are provided for training. Existing works [5, 17, 21, 26, 38] on semi-supervised 3D semantic segmentation have shown their effectiveness, but there is ample room for improvement in the following two aspects:

1) Utilization of language-vision models (LVM): Recently, LVM [16,33] trained on web-crawled image-text pairs has been applied in some downstream tasks [2, 31, 39, 50] to boost performance, owing to its strong representation ability. Text embeddings extracted by the LVM generally contain higher-level abstraction of visual concepts with rich semantic information, which is complementary to the 3D features that mainly contain geometric information and thus can facilitate mining the semantic information of unlabeled data. To our best knowledge, LVM has not been applied in semi-supervised 3D semantic segmentation yet.

2) Exploitation of unreliable points (unlabeled points with low-confidence predictions): The proportions of reliable and unreliable points predicted by the state-of-the-art semi-supervised 3D semantic segmentation method GPC [17] on ScanNet [4] dataset are illustrated in Fig. 1. As seen from this figure, a large number of unlabeled points are deemed unreliable, especially at the early training stage. Existing works [17,21] generally adopt a confidence-based filtering strategy to remove the unreliable points for guaranteeing the effectiveness and stability of the training process. However, simply discarding them will inevitably lose some useful information. Thus, a proper learning strategy is expected to be specially designed to exploit these unreliable points effectively.

Taking the above two aspects into consideration, we propose a method named **LASS3D** for semi-supervised 3D semantic segmentation, where the LVM and unreliable points are effectively exploited. The proposed LASS3D is built upon the MeanTeacher [37] framework, where the weights of the teacher branch are updated by the exponential moving average (EMA) of the student's weights.

In the student branch, two LVM (*i.e.*, KOSMOS-2 [32] and GroundedSAM [35]) are used to generate image-level and entity-level captions, and the images

are leveraged as the bridge to connect the text data and point clouds. Then, a semantic-aware adaptive fusion module (SAFM) is designed to assist point cloud learning with the semantic knowledge encoded in text data. Specifically, multi-level point-text pairs are constructed in SAFM to adaptively fuse the 3D features and the text embeddings of the image-level and entity-level captions, which could facilitate capturing both holistic and fine-grained semantic information.

In the teacher branch, a fusion adapter is utilized to narrow the distribution gap between the 3D features extracted by the backbone network and the text-enhanced 3D features in the student branch. Combined with the EMA strategy, the semantic information encoded in the text-enhanced 3D features is distilled from the student branch to the teacher branch. Note that we inject semantic information from text embeddings into the 3D features by fusion rather than directly mapping the semantic labels from pixels to points. Because the semantic labels generated by the open-world segmentation models (*e.g.*, SAM [20]) are generally in a one-hot form, which provide hard supervision. However, some generated synonymous labels may not exist in our desired label set (*e.g.*, the generated label is 'couch' but the ground-truth label is 'sofa'), and incorrect hard supervision may adversely affect the point cloud segmentation. Adaptively fusing the text embeddings and 3D features can mitigate this negative effect to some extent. Because the strong LVM can encode synonymous words into similar embeddings and thus can inject similar semantic information into 3D features. To effectively exploit the unlabeled points with unreliable predictions, a progressive exploitation strategy is designed based on the fact that determining the least likely category for an unlabeled sample is generally easier and more accurate than determining its most likely category. In addition, the proposed progressive exploitation strategy can be seamlessly integrated into some existing methods and further boost their performances, which is demonstrated in Sec. 4.4.

In summary, the contributions of this paper are as follows:

- We propose a semantic-aware adaptive fusion module (SAFM) to consolidate point cloud segmentation with the semantic information encoded in text embeddings. The multi-level fusion shows the feasibility and effectiveness of language-assisted 3D semantic segmentation in semi-supervised settings.
- We propose a progressive exploitation strategy for the unlabeled unreliable points. This strategy is beneficial to fully exploit the unlabeled data and can be seamlessly embedded into some existing methods.
- We propose the **LASS3D** for semi-supervised 3D semantic segmentation by integrating the above SAFM and progressive exploitation strategy. Its effectiveness is demonstrated by the results on both outdoor-scene and indoor-scene public datasets in Sec. 4.

2 Related Work

2.1 Semi-supervised 3D Semantic Segmentation

To alleviate the annotation burden, label-efficient 3D semantic segmentation (including weakly-supervised segmentation [13, 28, 29, 36] and semi-supervised seg-

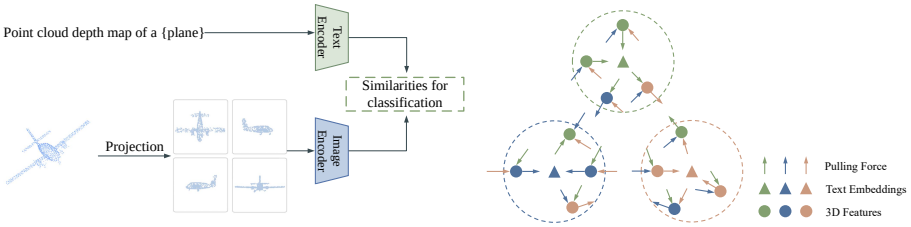
mentation [5, 17, 21, 25, 26, 38]) has drawn increasing interest among researchers. In this work, we delve into semi-supervised 3D semantic segmentation, which aims at utilizing a small number of densely labeled data and a large number of unlabeled data for model training. The core of semi-supervised 3D segmentation lies in exploiting the unlabeled data to the fullest. For example, Deng *et al*. [5] proposed to optimize the pseudo labels of the unlabeled data with the superpoints generated by geometry-based and color-based region-growing algorithms. Jiang *et al*. [17] proposed a label-guided contrastive loss for the unlabeled data, which could enhance the feature representation ability of the model. Kong *et al*. [21] proposed to leverage the spatial prior of LiDAR point clouds and use a mixing operation to provide supervisory information for the unlabeled data. Li *et al*. [26] proposed to utilize the reflectivity-prior descriptors to generate high-quality pseudo labels for the unlabeled data. Unal *et al*. [38] proposed to distill high-level feature information from a synthetically trained 2D network for data-efficient LiDAR semantic segmentation.

However, these methods neglect the usage of text data that contains rich semantic information. To our best knowledge, LASS3D is the first work to investigate the language-assisted semi-supervised 3D semantic segmentation, where the powerful representation ability of language-vision models is exploited.

2.2 Language-assisted 3D Learning

Large language-vision models (LVM) [16, 33] have drawn growing attention in computer vision recently, attributed to their powerful representation ability. For example, CLIP [33], which is trained on massive paired text-image data via contrastive learning, has prevailed in various downstream visual tasks, such as semantic segmentation [10, 18, 34, 48, 52], object detection [30, 45, 46], and video recognition [27]. Recent advanced methods that focused on language-assisted point cloud learning could be roughly divided into two types: the projection-based methods and the alignment-based methods.

The projection-based methods [9, 15, 49, 53] project the point clouds into the 2D plane so that the image encoder in the LVM can be seamlessly utilized and the text-image alignment of the LVM can be leveraged for downstream tasks.



(a) Pipeline of PointCLIP. **(b)** Alignment process of CLIP2Scene.

Fig. 2: Illustration of the pipeline of a projection-based method (*e.g*., PointCLIP [49]) and the alignment process of an alignment-based method (*e.g*., CLIP2Scene [2]).

Fig. 2a illustrates the general pipeline of PointCLIP [49], it classifies the point clouds according to the similarities between the text embeddings and features of the projected images. Nevertheless, the projection-based methods are generally designed for object-level 3D classification. In addition, they usually suffer from the loss of geometric information during the projection process, thus resulting in suboptimal performance.

The alignment-based methods [2,6,11,12,14,39,44,47] enforce the consistency among the multi-modal features. Fig. 2b illustrates the alignment process of CLIP2Scene [2], it uses contrastive learning to align the 3D features to their corresponding text embeddings. However, directly aligning multi-modal features may result in suboptimal results, due to the distribution discrepancy between the 3D features and text embeddings.

Unlike existing methods that are either projection-based or alignment-based, we propose a fusion-based method to adaptively fuse the embeddings of multi-level captions and 3D features, which can alleviate the loss of geometric information and mitigate the negative effect brought by the distribution gap.

3 Methodology

3.1 Architecture

The architecture of the proposed LASS3D is illustrated in Fig. 3. It is built upon the MeanTeacher [37] framework where the teacher branch provides the supervisory signals for the student branch and is updated by the exponential mean average (EMA) of the student's weights:

$$\theta_{teacher}^{t+1} = \alpha\theta_{teacher}^{t} + (1-\alpha)\theta_{student}^{t}, \tag{1}$$

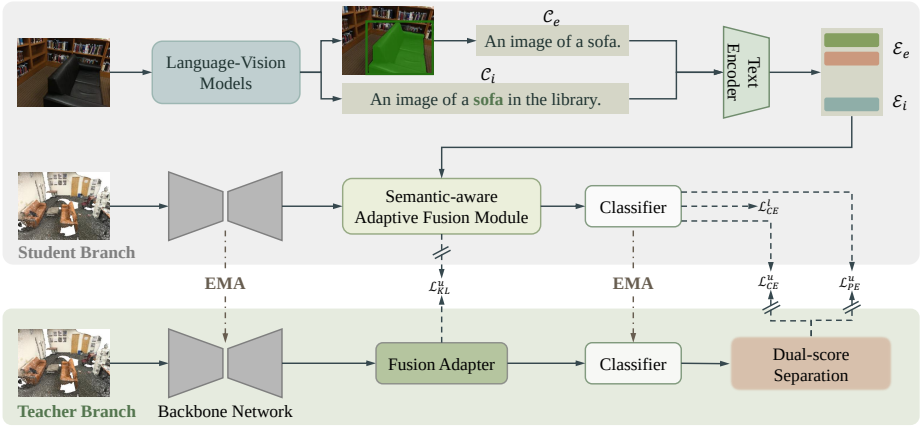


Fig. 3: Architecture of the proposed LASS3D. $\mathcal{C}_i$ and $\mathcal{C}_e$ stand for the image-level captions and entity-level captions. $\mathcal{E}_i$ and $\mathcal{E}_e$ are the embeddings of $\mathcal{C}_i$ and $\mathcal{C}_e$ respectively. '//' denotes the detach operation which is used to stop the gradient.

where $\theta^t_{teacher}$ and $\theta^t_{student}$ denote the weights of the teacher branch and student branch at time step t, and α is the update hyperparameter.

The student branch takes the point clouds and their corresponding images as input. The images are fed into the LVM, (*i.e.*, KOSMOS-2 [35] and GroundedSAM [32]) to produce image-level $\mathcal{C}_i$ and entity-level captions $\mathcal{C}_e$. Then, the multi-level captions are tokenized by the text encoder of CLIP [33] to generate their corresponding text embeddings. The text embeddings and 3D features extracted by the 3D backbone network are fused in the semantic-aware adaptive fusion module to obtain the text-enhanced 3D features, which are then used for segmentation in the student branch.

The teacher branch takes the same point clouds as input. The 3D features extracted by the 3D backbone network are aligned with the text-enhanced 3D features through a fusion adapter. Then, the classifier takes the aligned 3D features as input and outputs their corresponding prediction scores. A dual-score separation strategy is adopted to select reliable points and unreliable points from the unlabeled points. The predictions of the reliable points are regarded as the pseudo labels of their corresponding points in the student branch, where the cross-entropy loss $\mathcal{L}^u_{CE}$ is calculated. The predictions of the unreliable points in the student branch are used for the progressive exploitation, where the negative learning loss $\mathcal{L}^u_{PE}$ is calculated. Besides, the predictions of the labeled points in the student branch and their ground-truth labels are used to calculate the cross-entropy loss $\mathcal{L}^l_{CE}$, and the features output by the fusion adapter are aligned to the text-enhanced 3D features by a Kullback-Leibler divergence loss $\mathcal{L}^u_{KL}$.

Note that the text data is only required in the training stage. In the inference stage, only the teacher branch is used and it takes the point clouds as input.

3.2 Semantic-aware Adaptive Fusion Module

The architecture of the semantic-aware adaptive fusion module (SAFM) is illustrated in Fig. 4, it takes the embeddings of multi-level captions (*i.e.*, image-level embeddings $\mathcal{E}_i$ and entity-level embeddings $\mathcal{E}_e$) and 3D features f_{3D} as input and outputs the text-enhanced 3D features f^t_{3D}, which are then fed into the classifier for segmentation and used as the supervision for knowledge distillation.

The proposed SAFM is used to inject the semantic information encoded in multi-level captions into the 3D features. Specifically, the image-level captions $\mathcal{C}_i$ and entity-level captions $\mathcal{C}_e$ are generated by KOSMOS-2 [32] and GroundedSAM [35] respectively, and their corresponding embeddings are extracted by the text encoder of CLIP [33]. To guarantee the accuracy of multi-modal fusion, we construct the point-text pairs by using the images as the bridge. Then, given the point-text pairs, their corresponding features are adaptively fused to produce the text-enhanced 3D features. In this subsection, we will introduce the above two key operations of SAFM in detail.

Point-text pair construction. We use the images as the bridge to connect the text data and point cloud data. Specifically, the multi-level pixel-text pairs can be obtained by the language-vision models, *i.e.*, the image-level captions

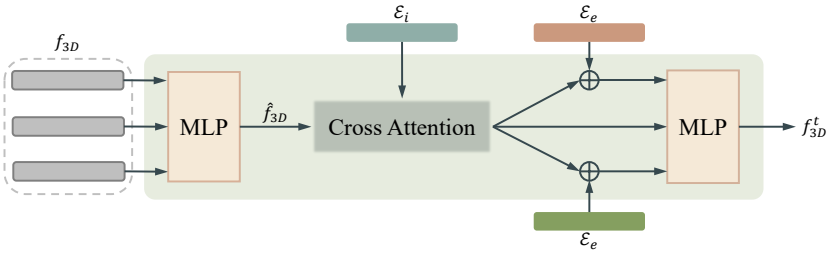


Fig. 4: Architecture of the proposed semantic-aware adaptive fusion module. f_{3D}, $\hat{f}_{3D}$, and f^t_{3D} denote the extracted 3D features, aligned 3D features, and text-enhanced 3D features. $\mathcal{E}_i$ and $\mathcal{E}_e$ are the embeddings of image-level and entity-level captions. $\oplus$ denotes the add operation.

correspond to all pixels of the given images, and the entity-level captions correspond to the pixels of the segmented regions. The pixel-point pairs can be easily obtained with the extrinsic and intrinsic parameters of the cameras:

$$[u, v, 1]^{\mathrm{T}} = \mathbf{KM}[x, y, z, 1]^{\mathrm{T}}, \tag{2}$$

where $[u, v, 1]^{\mathrm{T}}$ represents the homogeneous coordinates of a pixel, $[x, y, z, 1]^{\mathrm{T}}$ represents its corresponding 3D point position in world coordinates, $\mathbf{K}$ and $\mathbf{M}$ are the intrinsic and extrinsic parameters of the cameras. With the multi-level pixel-text and pixel-point pairs, the multi-level point-text pairs can be obtained.

Adaptive fusion. To effectively utilize the multi-level point-text pairs, an adaptive fusion strategy is proposed. As shown in Fig. 4, the dimension of the 3D features f_{3D} is firstly expanded by an MLP layer to align the embedding space of the text encoder, given that the text encoder in LASS3D is fixed. Since the image-level captions may have different correlations with different regions of the paired points, cross attention is performed between the aligned 3D features $\hat{f}_{3D}$ and image-level embeddings $\mathcal{E}_i$, with $\hat{f}_{3D}$ serving as the query vectors and $\mathcal{E}_i$ serving as the key vectors and value vectors. Then, considering that the points related to the same entity-level captions share the same semantic information, we simply add the entity-level embeddings $\mathcal{E}_e$ with their corresponding 3D features. Finally, an MLP layer is used to map the fused features to their original dimension and output the text-enhanced 3D features f^t_{3D}.

As shown in Fig. 3, the text-enhanced 3D features f^t_{3D} are not only fed into the classifier of the student branch for segmentation but also serve as the supervisory signals for knowledge distillation in the teacher branch. The features f_{ada} output by the fusion adapter in the teacher branch are aligned to their corresponding f^t_{3D} through $\mathcal{L}^u_{KL}$:

$$\mathcal{L}^u_{KL} = \mathrm{KL}(f^t_{3D} || f_{ada}), \tag{3}$$

where $\mathrm{KL}(\cdot)$ denotes the Kullback-Leibler divergence loss.

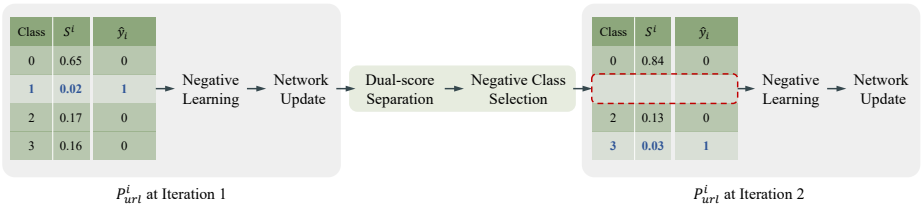


Fig. 5: Illustration of the progressive exploitation for the unreliable point P^i_{url} at the first two iterations. S^i denotes the prediction confidence of P^i_{url} and $\hat{y}_i$ denotes its negative label. The selected negative class (blue font) is used for negative learning and the used negative class (marked with red dashed box) is discarded in the subsequent iterations. Note that some unreliable points may turn into reliable points during progressive exploitation.

3.3 Progressive Exploitation Strategy

The progressive exploitation strategy is proposed to mine the useful information encapsulated in the unlabeled unreliable points. It contains two key steps: dual-score separation and progressive negative learning. In this subsection, we will introduce the above two steps in detail.

Dual-score separation. Unlike GPC [17] and LaserMix [21] which only use the confidences of predictions to measure the reliability, we propose a dual-score separation strategy that additionally introduces the variances among the predictions of different branches into reliability measurement. Specifically, the unlabeled points P_u in the teacher branch are separated into the reliable set P_{rl} and unreliable set P_{url} based on both confidences and variances:

$$\begin{aligned} &P_{rl} = B \cdot P_u,\ P_{url} = (1 - B) \cdot P_u, \\ &B = \mathbb{1}\Big[\sum_{c=1}^{C} \mathbb{1}[S_c \geq \tau_r] \cdot \mathbb{1}[\sigma_c \leq \tau_v] > 0\Big], \end{aligned} \tag{4}$$

where B is a binary mask to select P_{rl} and P_{url}, $\mathbb{1}$ is the indicator function, C denotes the number of classes, S_c denotes the prediction confidences of class c in the teacher branch, σ_c denotes the prediction variances of class c within the teacher and student branches, τ_r and τ_v are two pre-determined thresholds.

The predictions of reliable points are regarded as the pseudo labels of their corresponding points in the student branch, which are used for supervised training, as shown in Eq. (8). The unreliable points are fully exploited via progressive negative learning, which is elaborated below.

Progressive negative learning. Enlightened by the core idea of conventional negative learning [19, 40], we choose to mine information from the classes that the unreliable points do not belong to, rather than struggling to correctly tell which class they belong to. However, conventional methods only choose one negative class (the class with the lowest prediction score) for negative learning, which provides limited information and may cause unstable training due to the

dynamic negative labels. Instead, we conduct negative learning in a progressive manner, with more negative classes selected.

The progressive exploitation at the first two iterations is illustrated in Fig. 5. As shown in this figure, we maintain a list for each unreliable point to record its selected negative classes. In each iteration, one negative class is selected for each unreliable point:

$$\hat{y}_i[k] = \begin{cases} 1, & \text{if } k = \arg\min(S^i) \text{ and } S^i[k] \leq \tau_n \\ 0, & \text{otherwise} \end{cases}, \tag{5}$$

where $\hat{y}_i[k]$ denotes the k-th dimension of the negative label of P^i_{url}, S^i denotes the prediction confidence of P^i_{url}, and τ_n is a pre-determined threshold.

The selected negative label is used for negative learning:

$$\mathcal{L}^u_{PE} = -\sum_{i=1}^{N_u} \hat{y}_i \log(1 - O_i), \tag{6}$$

where N_u is the number of unreliable points with selected negative class and O_i is the prediction of P^i_{url}. Then, the used negative class is discarded in the subsequent iterations. Simultaneously, the network is optimized via the negative learning loss. Subsequently, a new negative class is selected from the remaining classes for the next iteration. The above process is repeated until no negative class can be selected or the iteration number is reached.

Moreover, it is worth noting that the proposed progressive exploitation strategy can be seamlessly integrated into methods that discard the unreliable points, and further boost their performances, as shown in Sec. 4.4.

3.4 Total Loss

The labeled points in the student branch are trained in a supervised manner, with their corresponding ground-truth labels y^l being the supervisory signals:

$$\mathcal{L}^l_{CE} = \mathrm{CE}(O^{stu,l}, y^l), \tag{7}$$

where $\mathrm{CE}(\cdot)$ denotes the cross-entropy loss, $O^{stu,l}$ denotes the predictions of the labeled points in the student branch.

The predictions of the reliable points in the teacher branch are regarded as the pseudo labels of their corresponding points in the student branch:

$$\mathcal{L}^u_{CE} = \mathrm{CE}(O^{stu,u}, y^u), \tag{8}$$

where y^u denotes the reliable predictions output by the teacher branch and $O^{stu,u}$ denotes the predictions of corresponding unlabeled points in the student branch.

Accordingly, the total loss $\mathcal{L}_{total}$ is the weighted sum of the two loss terms $\mathcal{L}^l_{CE}$, $\mathcal{L}^u_{CE}$ in the student branch and the aforementioned $\mathcal{L}^u_{PE}$, $\mathcal{L}^u_{KL}$:

$$\mathcal{L}_{total} = \mathcal{L}^l_{CE} + \lambda_u \mathcal{L}^u_{CE} + \lambda_{PE} \mathcal{L}^u_{PE} + \lambda_{KL} \mathcal{L}^u_{KL}, \tag{9}$$

where $\lambda_u, \lambda_{PE}, \lambda_{KL}$ are the weights of $\mathcal{L}^u_{CE}, \mathcal{L}^u_{PE}, \mathcal{L}^u_{KL}$.

4 Experiment

4.1 Datasets and Evaluation Metric

The following outdoor-scene and indoor-scene datasets are used to evaluate the proposed LASS3D:

- **SemanticKITTI** [1] is a large-scale 3D outdoor-scene LiDAR dataset consisting of 22 sequences, among which 10 sequences are used for training, 1 sequence is used for validation, and 11 sequences are used for testing. It contains 19 categories after merging and ignoring classes with few points. According to the official splitting, 850 scenes are used for training and validation, and 150 scenes are utilized for testing.
- **ScanNet V2** [4] is a widely-used 3D indoor-scene dataset consisting of 1613 scans with 20 categories, where the point clouds are reconstructed from multi-view RGB-D images. The training set contains 1201 scans, the validation split contains 312 scans, and the testing set contains 100 scans.

As done in previous works [21, 26, 38], we use the mIoU (mean Intersection over Union) as the evaluation metric.

4.2 Implementation Details

We follow the basic experimental settings of LaserMix [21] and GPC [17] for evaluating the proposed LASS3D on the outdoor-scene and indoor-scene datasets respectively. The classifiers and fusion adapter are both multi-layer perceptrons. We follow the settings of IGNet [38], which also adopts the MeanTeacher [37] framework, to set the update hyperparameter α as 0.999. The iteration number in progressive exploitation is set as 5. The hyperparameters $\tau_r, \tau_v, \tau_n, \lambda_u, \lambda_{PE}, \lambda_{KL}$ are set as 0.9, 0.05, 0.2, 0.1, 0.1, and 0.1 respectively.

Table 1: Comparative results on the SemanticKITTI [1] and ScanNet [4] datasets with varying labeled ratios. $\mathcal{P}$ denotes the point cloud data, $\mathcal{I}$ denotes the image data, and $\mathcal{T}$ denotes the text data. All mIoU scores are given in percentage (%). The best results are in **bold** and the second best results are marked with underlines.

Method	Modality	SemanticKITTI [1]				ScanNet [4]				
		1%	10%	20%	50%	5%	10%	20%	30%	40%
MeanTeacher [37]	$\mathcal{P}$	45.4	57.1	59.2	60.0	48.7	57.2	66.1	67.3	69.1
CBST [55]	$\mathcal{P}$	48.8	58.3	59.4	59.7	48.5	57.8	65.4	67.3	69.2
CPS [3]	$\mathcal{P}$	46.7	58.7	59.6	60.5	51.7	58.6	66.4	68.0	70.5
SSS-Net [5]	$\mathcal{P}$	-	-	-	-	-	52.4	55.1	-	-
LaserMix (Range View) [21]	$\mathcal{P}$	43.4	58.8	59.4	61.4	-	-	-	-	-
LaserMix (Voxel) [21]	$\mathcal{P}$	50.6	60.0	61.9	62.3	-	-	-	-	-
IGNet [38]	$\mathcal{P}, \mathcal{I}$	49.0	61.3	63.1	**64.8**	-	-	-	-	-
GPC [17]	$\mathcal{P}$	54.1	62.0	62.5	62.8	54.8	60.5	66.7	68.9	71.3
LiM3D [26]	$\mathcal{P}$	58.4	62.2	63.1	63.6	-	-	-	-	-
LASS3D	$\mathcal{P}, \mathcal{I}, \mathcal{T}$	**58.5**	**63.0**	**64.1**	64.5	**56.6**	**63.1**	**67.4**	**70.4**	**72.0**

4.3 Comparative Evaluation

We evaluate the proposed LASS3D on both outdoor-scene (SemanticKITTI [1]) and indoor-scene (ScanNet [4]) datasets in comparison to some classic semi-supervised methods [3, 37, 55] extended from the 2D domain and the state-of-the-art (SOTA) methods [5, 17, 21, 26, 38] that are specially designed for semi-supervised 3D semantic segmentation. Note that IGNet [38], LaserMix [21], and LiM3D [26] utilize some specific characters of the LiDAR point clouds, and thus cannot be applied to the ScanNet dataset. In addition, SSS-Net [5] leverages the color information of the point clouds, and thus cannot be applied to the SemantiKITTI dataset where the color information is unavailable.

Following the settings in [17, 21, 38], we set the labeled ratio of the outdoor-scene dataset and indoor-scene dataset as {1%, 10%, 20%, 40%} and {5%, 10%, 20%, 30%, 50%} respectively for a fair comparison. The comparative results are reported in Tab. 1. As seen from this table, the proposed LASS3D outperforms all the comparative methods that use uni-modal data as input, which demonstrates the effectiveness of LASS3D. In addition, LASS3D outperforms IGNet which takes both point clouds and images as input in most cases, which indicates that the proposed language-assisted method can better boost the 3D segmentation performances and may provide some insights on the exploitation of multi-modal data in semi-supervised 3D semantic segmentation.

Moreover, the segmentation results of LASS3D and the second-best methods are visualized in Fig. 6. As seen from this figure, LASS3D outperforms its second-best counterparts, demonstrating the effectiveness of LASS3D from the qualitative perspective.

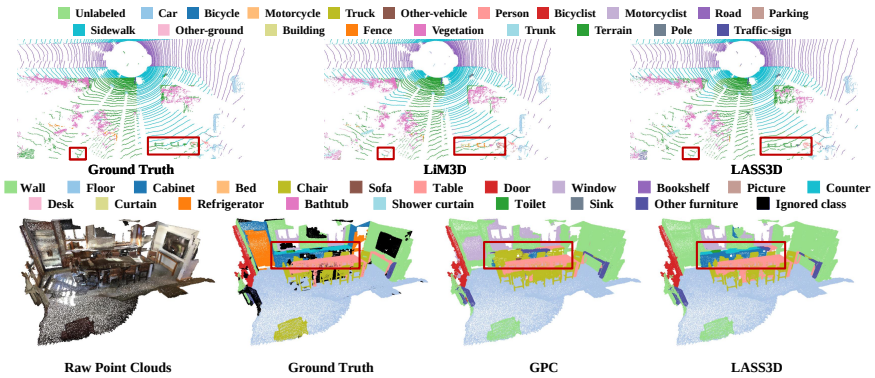


Fig. 6: Qualitative results on the SemanticKITTI [1] (top) and ScanNet [4] (bottom) datasets by LASS3D and the second-best methods. All models are trained with 10% labeled data. We use the red boxes to highlight the areas where LASS3D evidently outperforms the comparative methods.

Table 2: Ablation of the loss terms.

$\mathcal{L}_{CE}^{l}$	$\mathcal{L}_{CE}^{u}$	$\mathcal{L}_{KL}^{u}$	$\mathcal{L}_{PE}^{u}$	1%	10%	20%	50%
✓				49.6	58.8	60.7	61.2
✓	✓			53.7	60.1	61.6	62.8
✓	✓	✓		55.6	61.4	62.8	63.7
✓	✓	✓	✓	**58.5**	**63.0**	**64.1**	**64.5**

Table 3: Ablation of captions.

$\mathcal{C}_i$	$\mathcal{C}_e$	1%	10%	20%	50%
		50.0	59.8	61.5	62.1
✓		53.6	60.5	61.9	63.0
	✓	54.1	61.0	62.7	63.8
✓	✓	**58.5**	**63.0**	**64.1**	**64.5**

4.4 Ablation Study

We conduct ablation studies on SemanticKITTI [1] to investigate the effect of the key elements in the proposed LASS3D.

Effect of the loss terms. We investigate the effect of the loss terms and report the corresponding results in Tab. 2. As seen from the first two rows of Tab. 2, incorporating the unlabeled data into training can improve the performances. Then, as seen in the third row of Tab. 2, the $\mathcal{L}_{KL}^{u}$ is introduced into training and the performances are improved, indicating that the semantic information distilled from the text-enhanced 3D features is helpful for segmentation. Finally, in the last row of Tab. 2, the progressive exploitation is adopted, and the performances are further boosted, demonstrating that the proposed progressive exploitation can make the unreliable points beneficial to the model training.

Effect of different captions. We investigate the effect of different captions and report the corresponding results in Tab. 3. As seen from this table, both the image-level captions $\mathcal{C}_i$ and entity-level captions $\mathcal{C}_e$ can improve the segmentation performances, indicating that the semantic information encoded in text embeddings can facilitate the point cloud segmentation. When combining $\mathcal{C}_i$ and $\mathcal{C}_e$, the best performance is achieved. Because the $\mathcal{C}_i$ provides more holistic semantic information and $\mathcal{C}_e$ provides more fine-grained semantic information, which are complementary to each other. Thus, we utilize $\mathcal{C}_i$ and $\mathcal{C}_e$ in LASS3D.

Effect of the fusion strategy for the image-level embeddings $\mathcal{E}_i$. As mentioned in Sec. 3.2, we adopt the cross-attention operation to fuse $\mathcal{E}_i$ with the 3D features. To investigate the effect of the fusion strategy, we replace the cross-attention operation with add and concatenation operations. The corresponding results are reported in Tab. 4. As seen from this table, the model trained with the cross-attention operation achieves the best performance. Because the image-level captions usually contain holistic semantic information and

Table 4: Ablation of fusion strategy.

Strategy	1%	10%	20%	50%
Add	54.0	61.2	62.3	63.5
Concatenation	56.2	62.0	63.4	63.9
Cross attention	**58.5**	**63.0**	**64.1**	**64.5**

Table 5: Ablation of separation strategy.

Strategy	1%	10%	20%	50%
Confidence-based	57.7	62.3	63.5	64.1
Variance-based	57.0	62.1	62.8	63.5
Ours	**58.5**	**63.0**	**64.1**	**64.5**

may have different correlations with different regions of the paired points. Simply adding or concatenating $\mathcal{E}_i$ to all the paired 3D features leads to the lack of discrimination, while 3D semantic segmentation is a point-level prediction task that requires discriminative representations for each point. Cross attention can adaptively fuse $\mathcal{E}_i$ with 3D features and produce more discriminative features, which makes it a more suitable fusion strategy for $\mathcal{E}_i$.

Effectiveness of the semantic-aware adaptive fusion module (SAFM). As mentioned in Sec. 2, existing language-assisted point cloud learning methods are either projection-based or alignment-based. Project-based methods are generally designed for object-level 3D classification and are not applicable to scene-level 3D semantic segmentation. To verify the effectiveness of the proposed SAFM, we replace it with an alignment-based method (*e.g.*, the semantic consistency regularization in CLIP2Scene [2]). The corresponding results are reported in Tab. 6. As seen from this table, the model trained with SAFM outperforms the model trained with semantic consistency regularization. Probably because a large distribution discrepancy exists between the 3D features and text embeddings, simply aligning the 3D features to their corresponding text embeddings may result in the loss of some inherent geometric information. SAFM remedies this defect by adaptive fusion, and thus achieves better performance.

Effect of the separation strategy. As mentioned in Sec. 3.3, we adopt both confidences and variances to measure the reliability of the predictions and separate the unlabeled points into the reliable set and unreliable set. Here we evaluate the effectiveness of our separation strategy. The corresponding results are reported in Tab. 5. As seen from this table, the model trained with our strategy achieves the best performance. Probably because the confidence-based strategy alone is unable to filter out the highly confident wrong predictions, which may confuse the training. The variance-based strategy separates points according to the prediction consistency in different branches, which could be complementary to the confidence-based strategy in measuring reliability. Thus, we combine the above two strategies to measure reliability in a more comprehensive way, which can facilitate the model training.

Effect of the iteration number. The iteration number of progressive exploitation determines the utilization degree of the unreliable data. Here, we evaluate LASS3D with the iteration number set as $\{0, 1, 2, 3, 5, 7, 9, 11\}$, and the corresponding results are shown in Fig. 7. As seen from this figure, the performances are generally improved with the increase of iteration number. Because a larger iteration number means that the exploitation of the unreliable data is more thor-

Table 6: Ablation of semantic-aware adaptive fusion module (SAFM).

Strategy	1%	10%	20%	50%
Alignment-based	53.0	60.1	61.8	62.2
SAFM	**58.5**	**63.0**	**64.1**	**64.5**

Table 7: Ablation of the proposed progressive exploitation (PE).

Model	1%	10%	20%	50%
GPC [17]	54.1	62.0	62.5	62.8
GPC with PE	**54.7**	**62.8**	**63.1**	**63.5**

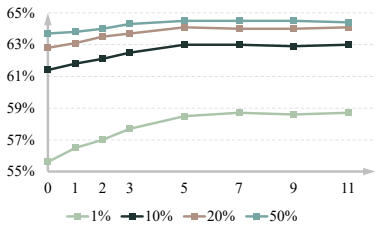


Fig. 7: Ablation of the iteration number. Note that the iteration here denotes the negative learning iteration in progressive exploitation, rather than the training iteration.

ough and thus brings better performance. However, when the iteration number reaches a certain level, the segmentation performance improves slightly. Probably because many unreliable points turn into reliable points during the progressive exploitation process, with most of their confident negative classes discarded. In addition, more iterations inevitably bring more training costs. Thus, we set the iteration number as 5 for a trade-off between the performance and training cost.

Effectiveness of the progressive exploitation. Considering that the proposed progressive exploitation strategy only requires the prediction confidences and variances, thus it can be seamlessly integrated into some existing methods. We integrate the progressive exploitation strategy into GPC [17] and report the comparative results in Tab. 7. As seen from this table, the proposed progressive exploitation strategy can further boost the performance of GPC, demonstrating the effectiveness of the proposed strategy for exploiting unreliable points.

4.5 Limitation

The proposed LASS3D has demonstrated its effectiveness in segmentation, however, it still has limitations. For example, LASS3D uses images as the bridge to connect the text data and point clouds, which is a complex operation and may bring accumulated errors in point-text matching. A more direct and accurate method to construct the point-text pairs is in demand.

5 Conclusion

In this paper, we propose a language-assisted semi-supervised 3D semantic segmentation method named LASS3D. In LASS3D, a semantic-aware adaptive fusion module is explored to fuse the 3D features with the embeddings of multi-level captions generated by language-vision models, which can inject the semantic information into the 3D features and thus facilitate 3D segmentation. In addition, a progressive exploitation strategy is explored for unlabeled unreliable data, which can further boost performance and can be seamlessly embedded into other methods. Experimental results on outdoor and indoor datasets show that LASS3D outperforms the comparative methods in most cases.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (Grant Nos. 61991423, 62376269, 62073199); the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA27040811); the Open Project Program of Key Laboratory of Industrial Internet and Big Data, China National Light Industry, Beijing Technology and Business University.

References

1. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: ICCV. pp. 9297–9307 (2019)
2. Chen, R., Liu, Y., Kong, L., Zhu, X., Ma, Y., Li, Y., Hou, Y., Qiao, Y., Wang, W.: Clip2scene: Towards label-efficient 3d scene understanding by clip. In: CVPR. pp. 7020–7030 (2023)
3. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: CVPR. pp. 2613–2622 (2021)
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR. pp. 5828–5839 (2017)
5. Deng, S., Dong, Q., Liu, B., Hu, Z.: Superpoint-guided semi-supervised semantic segmentation of 3d point clouds. In: ICRA. pp. 9214–9220 (2022)
6. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: Pla: Language-driven open-vocabulary 3d scene understanding. In: CVPR. pp. 7010–7019 (2023)
7. Fan, S., Dong, Q., Zhu, F., Lv, Y., Ye, P., Wang, F.: Scf-net: Learning spatial contextual features for large-scale point cloud segmentation. In: CVPR. pp. 14499–14508 (2021)
8. Graham, B., Engelcke, M., van der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. CVPR pp. 9224–9232 (2018)
9. Guo, Z., Zhang, R., Qiu, L., Ma, X., Miao, X., He, X., Cui, B.: Calip: Zero-shot enhancement of clip with parameter-free attention. In: AAAI. vol. 37, pp. 746–754 (2023)
10. He, W., Jamonnak, S., Gou, L., Ren, L.: Clip-s4: Language-guided self-supervised semantic segmentation. In: CVPR. pp. 11207–11216 (2023)
11. Hegde, D., Valanarasu, J.M.J., Patel, V.: Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In: ICCVW. pp. 2028–2038 (2023)
12. Hess, G., Tonderski, A., Petersson, C., Åström, K., Svensson, L.: Lidarclip or: How i learned to talk to point clouds. In: WACV. pp. 7438–7447 (2024)
13. Hu, Q., Yang, B., Fang, G., Guo, Y., Leonardis, A., Trigoni, N., Markham, A.: Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds. In: ECCV. pp. 600–619. Springer (2022)
14. Huang, R., Pan, X., Zheng, H., Jiang, H., Xie, Z., Wu, C., Song, S., Huang, G.: Joint representation learning for text and 3d point cloud. PR **147**, 110086 (2024)
15. Huang, T., Dong, B., Yang, Y., Huang, X., Lau, R.W., Ouyang, W., Zuo, W.: Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In: ICCV. pp. 22157–22167 (2023)
16. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. pp. 4904–4916 (2021)

17. Jiang, L., Shi, S., Tian, Z., Lai, X., Liu, S., Fu, C.W., Jia, J.: Guided point contrastive learning for semi-supervised point cloud semantic segmentation. In: ICCV. pp. 6423–6432 (2021)
18. Jiao, S., Wei, Y., Wang, Y., Zhao, Y., Shi, H.: Learning mask-aware CLIP representations for zero-shot segmentation. In: NeurIPS (2023)
19. Kim, Y., Yim, J., Yun, J., Kim, J.: Nlnl: Negative learning for noisy labels. In: ICCV. pp. 101–110 (2019)
20. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
21. Kong, L., Ren, J., Pan, L., Liu, Z.: Lasermix for semi-supervised lidar semantic segmentation. In: CVPR. pp. 21705–21715 (2023)
22. Lai, X., Chen, Y., Lu, F., Liu, J., Jia, J.: Spherical transformer for lidar-based 3d recognition. In: CVPR. pp. 17545–17555 (2023)
23. Lai, X., Liu, J., Jiang, L., Wang, L., Zhao, H., Liu, S., Qi, X., Jia, J.: Stratified transformer for 3d point cloud segmentation. In: CVPR. pp. 8500–8509 (2022)
24. Li, J., Dong, Q.: Open-set semantic segmentation for point clouds via adversarial prototype framework. In: CVPR. pp. 9425–9434 (2023)
25. Li, J., Dong, Q.: Density-guided semi-supervised 3d semantic segmentation with dual-space hardness sampling. In: CVPR. pp. 3260–3269 (2024)
26. Li, L., Shum, H.P., Breckon, T.P.: Less is more: Reducing task and model complexity for 3d point cloud semantic segmentation. In: CVPR. pp. 9361–9371 (2023)
27. Lin, Z., Geng, S., Zhang, R., Gao, P., de Melo, G., Wang, X., Dai, J., Qiao, Y., Li, H.: Frozen clip models are efficient video learners. In: ECCV. pp. 388–404. Springer (2022)
28. Liu, L., Zhuang, Z., Huang, S., Xiao, X., Xiang, T., Chen, C., Wang, J., Tan, M.: Cpcm: Contextual point cloud modeling for weakly-supervised point cloud semantic segmentation. In: ICCV. pp. 18413–18422 (2023)
29. Liu, Z., Qi, X., Fu, C.W.: One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. In: CVPR. pp. 1726–1736 (2021)
30. Lu, Y., Xu, C., Wei, X., Xie, X., Tomizuka, M., Keutzer, K., Zhang, S.: Open-vocabulary point-cloud object detection without 3d annotation. In: CVPR. pp. 1190–1199 (2023)
31. Ma, W., Li, S., Zhang, J., Liu, C.H., Kang, J., Wang, Y., Huang, G.: Borrowing knowledge from pre-trained language model: A new data-efficient visual learning paradigm. In: ICCV. pp. 18786–18797 (2023)
32. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
34. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: CVPR. pp. 18082–18091 (2022)
35. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded sam: Assembling open-world models for diverse visual tasks (2024)
36. Tang, L., Chen, Z., Zhao, S., Wang, C., Tao, D.: All points matter: entropy-regularized distribution alignment for weakly-supervised 3d segmentation. NeurIPS **36** (2024)

37. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. NIPS **30** (2017)
38. Unal, O., Dai, D., Hoyer, L., Can, Y.B., Van Gool, L.: 2d feature distillation for weakly-and semi-supervised 3d semantic segmentation. In: WACV. pp. 7336–7345 (2024)
39. Wang, Y., Huang, S., Gao, Y., Wang, Z., Wang, R., Sheng, K., Zhang, B., Liu, S.: Transferring clip's knowledge into zero-shot point cloud semantic segmentation. In: ACM MM. pp. 3745–3754 (2023)
40. Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., Le, X.: Semi-supervised semantic segmentation using unreliable pseudo-labels. In: CVPR. pp. 4248–4257 (2022)
41. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point transformer v3: Simpler, faster, stronger. In: CVPR (2024)
42. Wu, X., Lao, Y., Jiang, L., Liu, X., Zhao, H.: Point transformer v2: Grouped vector attention and partition-based pooling. In: NeurIPS. vol. 35, pp. 33330–33342 (2022)
43. Xiang, P., Wen, X., Liu, Y.S., Zhang, H., Fang, Y., Han, Z.: Retro-fpn: Retrospective feature pyramid network for point cloud semantic segmentation. In: ICCV. pp. 17826–17838 (2023)
44. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In: CVPR. pp. 1179–1189 (2023)
45. Yao, L., Han, J., Wen, Y., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, C., Xu, H.: Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. NeurIPS **35**, 9125–9138 (2022)
46. Yu, W., Liu, Y., Hua, W., Jiang, D., Ren, B., Bai, X.: Turning a clip model into a scene text detector. In: CVPR. pp. 6978–6988 (2023)
47. Zeng, Y., Jiang, C., Mao, J., Han, J., Ye, C., Huang, Q., Yeung, D.Y., Yang, Z., Liang, X., Xu, H.: Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In: CVPR. pp. 15244–15253 (2023)
48. Zhang, J., Fan, G., Wang, G., Su, Z., Ma, K., Yi, L.: Language-assisted 3d feature learning for semantic scene understanding. In: AAAI. pp. 3445–3453 (2023)
49. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: CVPR. pp. 8552–8562 (2022)
50. Zhang, Y., Huo, X., Chen, T., Wu, S., Wong, H.S.: Exploring intra-class variation factors with learnable cluster prompts for semi-supervised image synthesis. In: CVPR. pp. 7392–7401 (2023)
51. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: ICCV. pp. 16259–16268 (2021)
52. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: ECCV. pp. 696–712. Springer (2022)
53. Zhu, X., Zhang, R., He, B., Guo, Z., Zeng, Z., Qin, Z., Zhang, S., Gao, P.: Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In: ICCV. pp. 2639–2650 (2023)
54. Zhu, X., Zhou, H., Wang, T., Hong, F., Ma, Y., Li, W., Li, H., Lin, D.: Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In: CVPR. pp. 9939–9948 (2021)
55. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV. pp. 289–305 (2018)