Generalized Few-shot 3D Point Cloud Segmentation with Vision-Language Model

Zhaochong An[1,2], Guolei Sun[2*], Yun Liu[3*], Runjia Li[4], Junlin Han[4], Ender Konukoglu[2], Serge Belongie[1]
[1] Department of Computer Science, University of Copenhagen
[2] Computer Vision Laboratory, ETH Zurich
[3] College of Computer Science, Nankai University
[4] Department of Engineering Science, University of Oxford

Abstract

Generalized few-shot 3D point cloud segmentation (GFS-PCS) adapts models to new classes with few support samples while retaining base class segmentation. Existing GFS-PCS methods enhance prototypes via interacting with support or query features but remain limited by sparse knowledge from few-shot samples. Meanwhile, 3D vision-language models (3D VLMs), generalizing across open-world novel classes, contain rich but noisy novel class knowledge. In this work, we introduce a ***GFS****-PCS framework that synergizes dense but noisy pseudo-labels from 3D* ***VL****Ms with precise yet sparse few-shot samples to maximize the strengths of both, named* ***GFS-VL****. Specifically, we present a prototype-guided pseudo-label selection to filter low-quality regions, followed by an adaptive infilling strategy that combines knowledge from pseudo-label contexts and few-shot samples to adaptively label the filtered, unlabeled areas. Additionally, we design a novel-base mix strategy to embed few-shot samples into training scenes, preserving essential context for improved novel class learning. Moreover, recognizing the limited diversity in current GFS-PCS benchmarks, we introduce two challenging benchmarks with diverse novel classes for comprehensive generalization evaluation. Experiments validate the effectiveness of our framework across models and datasets. Our approach and benchmarks provide a solid foundation for advancing GFS-PCS in the real world. The code is at here.*

1. Introduction

Understanding dense 3D semantics is essential for many vision applications [20, 35, 42, 52, 53, 61, 73], and *few-shot point cloud semantic segmentation* (FS-PCS) has emerged as a valuable task [1, 77], enabling models to extend from base to novel classes with minimal annotations for novel classes. However, typical few-shot models require additional support samples for each novel class at inference and only predict novel classes, ignoring base classes. To address this, *generalized few-shot point cloud semantic segmentation* (GFS-PCS) [66] was introduced, allowing models to directly segment both base and novel classes after few-shot adaptation, making it more practical for real-world use.

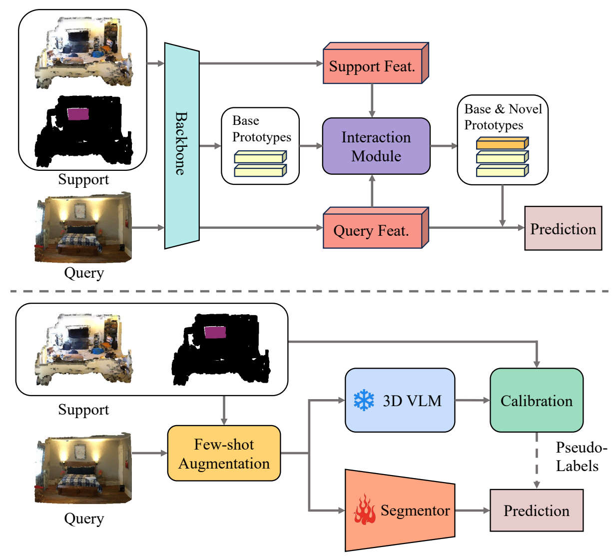


Figure 1. **Comparison of our framework with previous work.** *Top:* Prior work [56, 66] primarily enhances prototypes through interaction modules that integrate support/query features, making predictions based on refined prototypes. However, they are limited by the sparse knowledge from few-shot samples. *Bottom:* Our framework addresses this limitation by leveraging the extensive open-world knowledge from 3D VLMs through pseudo-labels. We mitigate the noise inherent in 3D VLMs by calibrating their raw pseudo-labels with precise few-shot samples, thereby effectively expanding novel class knowledge while ensuring reliability.

*Corresponding authors: Guolei Sun and Yun Liu

Current GFS-PCS models [17, 33, 34, 66] primarily utilize prototype learning [51], representing each class as a prototype and predicting based on the relationship between query points and these prototypes. As shown in Fig. 1, these methods mainly focus on refining prototypes through interactions with support/query samples for enhanced segmentation. For instance, CAPL [56] adapts base prototypes to novel classes using co-occurrence knowledge from support samples and contextual information from queries. GW [66] encodes shared geometric structures into geometry prototypes to enhance semantic prototypes. However, these approaches remain limited in novel class generalization due to the sparse knowledge available from few-shot samples.

In parallel, *3D vision-language models* (3D VLMs) have been developed to enable open-vocabulary recognition by aligning 3D and language features. Leveraging language models trained on vast open-text data, 3D VLMs exhibit strong generalization abilities, allowing recognition of open-set classes in 3D. Since paired 3D-text data is scarce, some approaches [5, 18, 44, 54, 75] distill 2D features from 2D VLMs [12, 27] into their 3D encoders, while others [9, 10, 21, 67] use captioning models [47, 60] to generate scene- or region-level descriptions, enabling point-language alignment for direct 3D learning. Recognizing the open-world potential of 3D VLMs, we propose utilizing their rich knowledge to enhance GFS-PCS. A straightforward method to integrate 3D VLMs [44, 67] is to generate dense pseudo-labels of novel classes as additional supervision. However, predictions from 3D VLMs are often noisy, compounding errors in GFS-PCS models. Meanwhile, sparse support samples offer accurate annotations for novel classes. Therefore, given the dense but noisy pseudo-labels from 3D **VL**Ms and the accurate yet limited support samples, we propose a new **GFS**-PCS framework, named **GFS-VL**, to combine the strengths of both, as in Fig. 1.

Specifically, GFS-VL incorporates three novel techniques. First, we introduce a pseudo-label selection technique that uses accurate few-shot data to filter pseudo-labels, retaining only high-quality regions while excluding noisy predictions. Second, as filtered wrong predictions will leave some regions unlabeled, potentially corresponding to novel objects, we present an adaptive infilling approach to enrich these regions. It combines knowledge from pseudo-label contexts and few-shot samples to construct an adaptive prototype set to label unlabeled regions, effectively considering both the completion of incomplete masks and the discovery of missing classes. Third, to further utilize few-shot samples, we propose a novel-base mix strategy, embedding support samples into training scenes. Unlike traditional 3D data augmentation [39, 49, 64, 65, 71, 82], which mainly aims at fully-supervised segmentation and mixes object contexts, our approach emphasizes preserving contextual cues, which is crucial for novel class learning [56] by helping identify challenging novel classes.

Furthermore, we identify limitations in current evaluation benchmarks. Existing benchmarks based on ScanNet [7] and S3DIS [3] datasets include only six novel classes, limiting diversity and failing to represent the complexity of real-world scenarios where novel classes are constantly varying. To address this, we introduce two challenging benchmarks: one with 40 novel classes from ScanNet200 [48] and another with 18 novel classes from ScanNet++ [70]. As detailed in Sec. 3.2, these benchmarks provide broader and more representative coverage of novel classes, enabling a more comprehensive evaluation of models' generalization capabilities.

By fully integrating the benefits of 3D VLMs and few-shot data, our approach achieves state-of-the-art GFS-PCS performance on both existing and newly established benchmarks. Experiments demonstrate the effectiveness and generalizability of our framework across various models and datasets. Additionally, previous baselines exhibit limited performance when evaluated on our new benchmarks, underscoring the necessity of our benchmarks for assessing real-world generalization. Together, our methods and benchmarks would offer critical insights and tools to advance future research on GFS-PCS.

2. Related Work

2.1. Few-shot 3D Point Cloud Segmentation

3D point cloud segmentation is fundamental in understanding scene semantics, with many fully-supervised methods advancing this field [8, 15, 23, 24, 28, 40, 43, 58, 63, 76, 78]. However, these methods need expensive large-scale point-level annotations and have fixed output spaces. To address these limitations, FS-PCS was introduced in attMPTI [77], aiming to generalize to novel classes with limited support samples. FS-PCS research can be divided into two categories based on how relationships between query points and support classes are modeled: i) Feature optimization [16, 31, 37, 41, 59, 62, 72, 80, 81] – These methods refine support prototypes or query features to enhance class separation. Final predictions are made using non-parametric, distance-based metrics, which implicitly model support-query relationships. ii) Correlation optimization [1, 2] – These approaches directly optimize correlations between support and query samples, explicitly modeling their relationships. COSeg [1] pioneered this approach recently and corrected two issues of foreground leakage and sparse point distribution in the previous FS-PCS setting.

2.2. Generalized Few-shot 3D Point Cloud Segmentation

While standard few-shot models adapt effectively to novel classes with limited data, they are constrained to predict only novel classes and require support samples to specify target classes during inference. A more practical task, generalized few-shot segmentation, occurs to require predicting both base and novel classes at inference without support samples, as first introduced in 2D segmentation [4, 14, 17, 25, 33, 34, 36, 38, 56]. The pioneer work PIFS [4], using the prototype learning paradigm [51], fine-tunes base and novel prototypes with a distillation loss, while CAPL [56] refines base prototypes using co-occurrence priors from support samples and dynamic contextual information from queries. For 3D, GW [66] introduces this setting to point cloud segmentation. GW models shared local geometric structures across base and novel classes as "geometry words" and then builds geometric prototypes to enhance the semantic prototypes, which are learned similarly to CAPL [56] by leveraging contextual information. To mine background semantics, Tsai *et al*. [57] clustered background points to generate pseudo-class prototypes distinct from base classes, leveraging multiple 2D views and 2D foundation models to link these points with class prompts.

2.3. 3D Vision-Language Models

3D VLMs align 3D point cloud features with language features, enabling open-world 3D understanding. However, developing these models poses unique challenges compared to 2D, mainly due to the scarcity of paired 3D-text data. To address this, recent work leverages multi-view 2D images, commonly associated with 3D point clouds, as intermediaries [30]. Some methods [5, 18, 19, 44, 54, 60, 75, 79] distill 2D features from 2D VLMs [12, 27, 46, 68] into their 3D encoders. For instance, OpenScene [44] aligns 3D and text representations by optimizing 3D-2D alignment using 2D VLMs [12, 27]. However, extracting these 2D features is computationally expensive, and the learned 3D features may inherit 2D prediction errors. Other approaches [9, 10, 21] generate point-language paired data by using captioning models [47, 60] to produce text descriptions of images. While effective, these captions are often at the scene level, limiting their ability to capture fine-grained 3D features. RegionPLC [67] recently introduced high-quality region-level 3D-language associations, supporting robust 3D learning by dense regional language supervision.

2.4. Point Cloud Data Augmentation

To address data limitations in the 3D domain, numerous methods have been developed to expand point cloud distributions. One category augments individual point clouds by altering geometric properties [22, 29, 32, 45, 50], using techniques such as shape transformations and patch shuffling. Another approach uses mixing techniques for 3D objects [6, 26, 74] and 3D scenes [11, 49, 64, 65, 71, 82]. For instance, Mix3D [39] mixes points from two scenes as an out-of-context augmentation for semantic segmentation, while methods like [49] select domain-specific points with semantic information to mix across domains. Other methods, such as [64], relocate objects to less frequent locations to vary spatial distributions. Notably, most augmentation methods in semantic segmentation [39, 64] and detection [82] modify the original context of target classes, pushing models to learn object patterns independently of surroundings. However, we argue that preserving contextual dependencies is crucial for GFS-PCS, where novel classes often involve challenging, hard-to-detect objects. Isolated object patterns alone are insufficient for effective novel class generalization [56]. In contrast, our proposed novel-base mix augmentation retains key contextual information when incorporating support samples into training scenes, enhancing models' ability to recognize novel classes.

Dataset	Base	Novel	Max (F)	Min (F)	Max (P)	Min (P)
S3DIS	7	6	185	29	59,929	30,013
ScanNet	13	6	411	133	4,479	1,148
ScanNet200	12	45	**733**	**102**	12,641	279
ScanNet++	12	18	143	82	**84,375**	**604**

Table 1. Statistics for GFS-PCS benchmarks across four datasets. Base/Novel indicates the number of base and novel classes. Max/Min (F) is the maximum and minimum occurrences of each novel class across the entire dataset, while Max/Min (P) is the maximum and minimum average number of points per novel class.

3. GFS-PCS Overview

3.1. Problem Definition

GFS-PCS requires the model to identify both base and novel classes present in test scenes. Let $\boldsymbol{C}^{\mathrm{b}}$ denote the set of base class names with size N_b and $\boldsymbol{C}^{\mathrm{n}}$ the set of novel class names with size N_n. The total number of classes is $N_c = N_b + N_n$. The base and novel class sets are mutually exclusive, *i.e.*, $\boldsymbol{C}^{\mathrm{b}} \cap \boldsymbol{C}^{\mathrm{n}} = \varnothing$. Evaluation is conducted on the test dataset $\mathbf{D}_{\mathrm{test}}$ to assess segmentation across all classes $\boldsymbol{C}^{\mathrm{b}} \cup \boldsymbol{C}^{\mathrm{n}}$. For training, the model is first trained on base classes and then registered with novel classes. The base dataset $\mathbf{D}_{\mathrm{base}}$ is defined as $\mathbf{D}_{\mathrm{base}} = \{\mathbf{X}_{\mathrm{b}}^{\mathrm{i}}, \mathbf{Y}_{\mathrm{b}}^{\mathrm{i}}\}_i$, where $\mathbf{X}_{\mathrm{b}}^{\mathrm{i}}$ represents the i-th point cloud and $\mathbf{Y}_{\mathrm{b}}^{\mathrm{i}}$ contains the corresponding base class labels. For the novel class dataset, we define $\mathbf{D}_{\mathrm{novel}}$ as the collection of K-shot support samples: $\{\{\mathbf{X}_{\mathrm{k}}^{\mathrm{c}}, \mathbf{Y}_{\mathrm{k}}^{\mathrm{c}}\}_{k=1}^{K}\}_{c=N_b}^{N_c}$. Each novel class c has K support samples $\{\mathbf{X}_{\mathrm{k}}^{\mathrm{c}}\}_{k=1}^{K}$ with exclusive labels $\{\mathbf{Y}_{\mathrm{k}}^{\mathrm{c}}\}_{k=1}^{K}$. The base classes occupy class indices in the range $[0, N_b)$, while novel classes are indexed in the range $[N_b, N_c)$, with background regions labeled as -1. For simplicity, we denote a data sample from $\mathbf{D}_{\mathrm{base}}$ as $\mathbf{X}_{\mathrm{b}}$ and $\mathbf{Y}_{\mathrm{b}}$, while $\mathbf{X}_{\mathrm{n}}^{\mathrm{c}}$ and $\mathbf{Y}_{\mathrm{n}}^{\mathrm{c}}$ refer to a support sample of class c in $\mathbf{D}_{\mathrm{novel}}$.

3.2. New Evaluation Benchmarks

Current evaluation benchmarks for GFS-PCS utilize the ScanNet [7] and S3DIS [3] datasets. However, we iden-

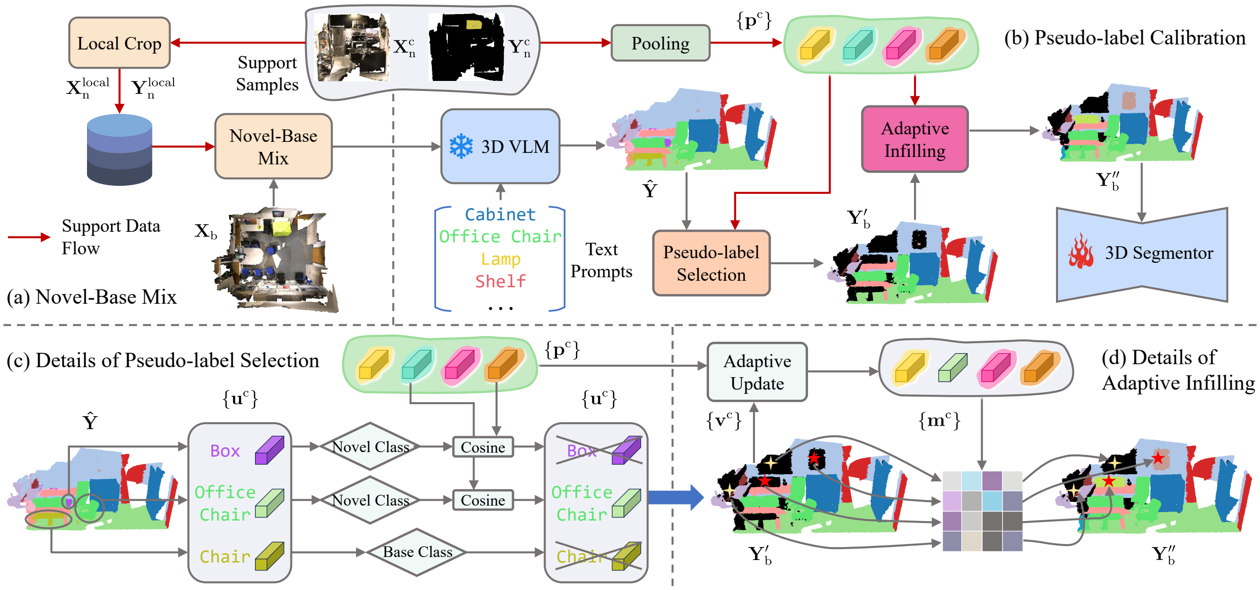


Figure 2. **Overview of the proposed GFS-VL.** (a), (b) Given an input point cloud $\mathbf{X}_\mathrm{b}$, we apply a novel-base mix to embed support samples into the training scene while preserving essential context. The scene is then processed by a 3D VLM, using all class names as prompts to generate raw predictions $\hat{\mathbf{Y}}$. Leveraging support prototypes $\{\mathbf{p}^\mathrm{c}\}$, the raw predictions undergo pseudo-label selection to filter out noisy regions, followed by adaptive infilling to label the filtered, unlabeled areas, yielding refined supervision $\mathbf{Y}''_\mathrm{b}$ for training the 3D segmentor. (c), (d) illustrate the details of the pseudo-label selection and adaptive infilling processes.

tify limitations in the number and diversity of novel classes in these benchmarks. As shown in Tab. 1, while ScanNet includes 13 base classes and S3DIS has 7, both datasets include only 6 novel classes. Our analysis of the frequency and the average number of points per novel class further highlights the lack of diversity among these limited novel classes in these two benchmarks. This indicates that current evaluations are not sufficiently representative of the complexity of real-world novel categories, failing to robustly assess models' generalization abilities. To address this, we propose two new and more challenging evaluation benchmarks based on ScanNet200 [48] and ScanNet++ [70] datasets. These new benchmarks, as detailed in Tab. 1, feature a larger number and greater diversity of novel classes, providing a more realistic and comprehensive testbed for evaluating model generalization to novel categories.

4. Method

4.1. Overview

In our proposed framework GFS-VL, we adopt a canonical segmentor architecture consisting of a backbone and a linear classification head. This minimalistic structure retains flexibility and simplicity, facilitating reproducibility.

Following the standard GFS-PCS procedure [66], the segmentor is initially trained on base classes. The backbone, denoted as Φ, and the linear classifier for base classes, $\mathcal{H}_b$, are employed to make base class predictions $\mathbf{P}_\mathrm{b}$:

$$\mathbf{F} = \Phi(\mathbf{X}_\mathrm{b}) \in \mathbb{R}^{N_p \times D}, \mathbf{P}_\mathrm{b} = \mathcal{H}_b(\mathbf{F}) \in \mathbb{R}^{N_p \times N_b}, \quad (1)$$

where N_p is the number of points in the point cloud input $\mathbf{X}_\mathrm{b}$, and $\mathbf{F}$ represents backbone features with a channel dimension of D. Then, when registering novel classes, a new linear classifier $\mathcal{H}_n$ is introduced to handle the novel classes. The concatenated predictions from both classifiers form the complete output $\mathbf{P}$ for all base and novel classes:

$$\mathbf{P} = [\mathcal{H}_b(\mathbf{F}), \mathcal{H}_n(\mathbf{F})] \in \mathbb{R}^{N_p \times N_c}, \quad (2)$$

where $[\cdot, \cdot]$ denotes concatenation. Subsequently, the model is fine-tuned on the base and few-shot samples, enabling it to simultaneously segment both base and novel classes. To fully exploit the rich knowledge from 3D VLMs while minimizing potential noise interference, we propose GFS-VL, as shown in Fig. 2, to maximize the utility of limited but accurate novel samples to guide the learning process. We detail each designed module in the following sections, presented under the 1-shot setting for clarity.

4.2. Pseudo-label Selection

A direct method for utilizing 3D VLMs in GFS-PCS is to use their predictions as pseudo-labels for novel classes. However, these raw predictions are noisy, hindering few-shot models from learning good novel class representations. Moreover, such noisy pseudo-labels could introduce error accumulation from the 3D VLMs into the few-shot models.

Therefore, we introduce an effective pseudo-label selection method by utilizing the valuable few-shot support samples to guide the selection of reliable novel class predictions. Specifically, for each novel class, we first compute support prototypes using the few-shot samples. This is achieved by applying the vision encoder Θ_v of the 3D VLM Θ to compute masked average features for each novel class:

$$\begin{aligned} &\mathbf{F}^{\mathrm{c}}_{\mathrm{n}} = \Theta_v(\mathbf{X}^{\mathrm{c}}_{\mathrm{n}}) \in \mathbb{R}^{N_p \times D_v}, \\ &\mathbf{p}^{\mathrm{c}} = \frac{\sum_{i=1}^{N_p} \mathbf{F}^{\mathrm{c}}_{\mathrm{n,i}} \mathbf{Y}^{\mathrm{c}}_{\mathrm{n,i}}}{\sum_{i=1}^{N_p} \mathbf{Y}^{\mathrm{c}}_{\mathrm{n,i}}}, \qquad c = N_b, \ldots, N_c - 1, \end{aligned} \tag{3}$$

where D_v is the feature dimension of the 3D VLM, and $\mathbf{p}^{\mathrm{c}} \in \mathbb{R}^{D_v}$ represents the support prototype for novel class c. For clarity, we define this prototype extraction process as $\mathcal{F}_{\mathrm{pool}}$. Thus, Eq. (3) becomes: $\mathbf{p}^{\mathrm{c}} = \mathcal{F}_{\mathrm{pool}}(\mathbf{X}^{\mathrm{c}}_{\mathrm{n}}, \mathbf{Y}^{\mathrm{c}}_{\mathrm{n}})$.

Next, given the current base training input $\mathbf{X}_{\mathrm{b}}$ with base labels $\mathbf{Y}_{\mathrm{b}}$, we prompt the 3D VLM Θ using all base and novel class names to obtain predictions $\hat{\mathbf{Y}} \in \mathbb{R}^{N_p}$. Let $\hat{\boldsymbol{C}}_{\mathrm{n}}$ be the novel class indices existing in $\hat{\mathbf{Y}}$. We then compute the predicted prototype $\mathbf{u}^{\mathrm{c}}$ for each novel class in $\hat{\boldsymbol{C}}_{\mathrm{n}}$:

$$\mathbf{u}^{\mathrm{c}} = \mathcal{F}_{\mathrm{pool}}(\mathbf{X}_{\mathrm{b}}, \mathbf{1}_{[\hat{\mathbf{Y}}=c]}), \qquad c \in \hat{\boldsymbol{C}}_{\mathrm{n}}. \tag{4}$$

Here, $\mathbf{1}_{[\hat{\mathbf{Y}}=c]}$ is a binary mask, set to 1 where $\hat{\mathbf{Y}}$ equals the class index c and to 0 otherwise. Then, we can filter the raw predictions to select high-quality novel class pseudo-labels:

$$\hat{\mathbf{Y}}_{\mathrm{i}} = \begin{cases} -1, & \text{if } \hat{\mathbf{Y}}_{\mathrm{i}} \in [0, N_b) \text{ or} \\ & \quad \big(\hat{\mathbf{Y}}_{\mathrm{i}} \in [N_b, N_c) \text{ and} \\ & \quad\quad \mathrm{sim}(\mathbf{u}^{\hat{\mathbf{Y}}_{\mathrm{i}}}, \mathbf{p}^{\hat{\mathbf{Y}}_{\mathrm{i}}}) < \tau\big), \\ \hat{\mathbf{Y}}_{\mathrm{i}}, & \text{otherwise.} \end{cases} \tag{5}$$

Here, if the predicted label $\hat{\mathbf{Y}}_{\mathrm{i}}$ for the i-th point is a base class, or a novel class with cosine similarity below a threshold τ between the predicted class prototype $\mathbf{u}^{\hat{\mathbf{Y}}_{\mathrm{i}}}$ and the support prototype $\mathbf{p}^{\hat{\mathbf{Y}}_{\mathrm{i}}}$, we filter this pseudo-label by setting it to -1. Otherwise, we retain the original pseudo-label. Note this filtering process can be efficiently implemented using mask-based indexing without iterating each point.

Now the updated $\hat{\mathbf{Y}}$ contains only reliable pseudo-labels for novel classes. Given the original base class labels $\mathbf{Y}_{\mathrm{b}}$, its background region (labeled as -1) serves as a potential area for novel classes. Therefore, we merge the updated $\hat{\mathbf{Y}}$ into the background region in $\mathbf{Y}_{\mathrm{b}}$ to generate augmented labels $\mathbf{Y}'_{\mathrm{b}}$ with additional reliable novel class pseudo-labels:

$$\begin{aligned} \mathbf{Y}'_{\mathrm{b}} &= \mathbf{Y}_{\mathrm{b}}, \\ \mathbf{Y}'_{\mathrm{b}}[\mathbf{Y}'_{\mathrm{b}} = -1] &= \hat{\mathbf{Y}}[\mathbf{Y}'_{\mathrm{b}} = -1]. \end{aligned} \tag{6}$$

4.3. Adaptive Infilling

After selection, $\mathbf{Y}'_{\mathrm{b}}$ includes reliable supervision for novel classes, while some regions remain unlabeled due to filtered low-quality predictions. These filtered predictions from the 3D VLM usually assign wrong labels, either by entirely missing true novel areas or partially mislabeling them [55]. Consequently, $\mathbf{Y}'_{\mathrm{b}}$ contains unlabeled regions that potentially correspond to missing or incomplete novel labels.

To address these gaps, we propose an adaptive infilling approach that utilizes both the few-shot samples and the current labels $\mathbf{Y}'_{\mathrm{b}}$ to build an adaptive prototype set for novel classes. This set allows us to assign novel labels adaptively to unlabeled regions, ensuring more comprehensive coverage of novel classes. We begin by extracting novel class prototypes from $\mathbf{Y}'_{\mathrm{b}}$:

$$\mathbf{v}^{\mathrm{c}} = \mathcal{F}_{\mathrm{pool}}(\mathbf{X}_{\mathrm{b}}, \mathbf{1}_{[\mathbf{Y}'_{\mathrm{b}}=c]}), \quad c \in \boldsymbol{C}^{\mathrm{y}}_{\mathrm{n}}, \tag{7}$$

where $\boldsymbol{C}^{\mathrm{y}}_{\mathrm{n}}$ represents the novel class indices present in $\mathbf{Y}'_{\mathrm{b}}$. Using both these extracted prototypes and pre-computed support prototypes from the few-shot samples, we construct an adaptive prototype set, defined as $\{\mathbf{m}^{\mathrm{c}}\}$, where:

$$\mathbf{m}^{\mathrm{c}} = \begin{cases} \mathbf{v}^{\mathrm{c}}, & \text{if } c \in \boldsymbol{C}^{\mathrm{y}}_{\mathrm{n}}, \\ \mathbf{p}^{\mathrm{c}}, & \text{otherwise,} \end{cases} \quad \text{for } c = N_b, \ldots, N_c - 1. \tag{8}$$

This set $\{\mathbf{m}^{\mathrm{c}}\}$ incorporates novel class prototypes $\mathbf{v}^{\mathrm{c}}$ from $\mathbf{Y}'_{\mathrm{b}}$ if they exist; otherwise, it defaults to the few-shot support prototypes $\mathbf{p}^{\mathrm{c}}$. By adapting to the current pseudo-labels, this set facilitates the completion of incomplete novel class pseudo-labels while allowing for the discovery of missed novel classes based on support prototypes.

Next, we initialize $\mathbf{Y}''_{\mathrm{b}} = \mathbf{Y}'_{\mathrm{b}}$. For each unlabeled point $\mathbf{Y}''_{\mathrm{b,i}}$ with feature $\mathbf{F}_{\mathrm{b,i}}$ from Θ_v, we calculate its cosine similarity with each prototype $\mathbf{m}^{\mathrm{c}}$ as $S^{\mathrm{c}}_{\mathrm{b,i}} = \mathrm{sim}(\mathbf{F}_{\mathrm{b,i}}, \mathbf{m}^{\mathrm{c}})$ and assign the corresponding novel class label if the maximum similarity exceeds a threshold δ:

$$\mathbf{Y}''_{\mathrm{b,i}} = \begin{cases} \arg\max_{c} S^{\mathrm{c}}_{\mathrm{b,i}}, & \text{if } \max_{c} S^{\mathrm{c}}_{\mathrm{b,i}} \geq \delta, \\ -1, & \text{otherwise.} \end{cases} \tag{9}$$

This adaptive infilling mechanism effectively integrates knowledge from few-shot support samples with the current pseudo-label context, creating adaptive prototypes that help discover missed novel objects and complete partial pseudo-labels, thereby enhancing the quality of novel region labels.

4.4. Novel-Base Mix

To more sufficiently utilize support samples, we introduce a novel-base mix approach that effectively integrates these valuable samples with the training data. Specifically, we start by randomly sampling a novel sample $\mathbf{X}^{\mathrm{c}}_{\mathrm{n}}$ from $\mathbf{D}_{\mathrm{novel}}$. To enhance the model's focus on novel class, we crop the local bounding region based on the novel class mask $\mathbf{Y}^{\mathrm{c}}_{\mathrm{n}}$:

$$\mathbf{X}^{\mathrm{local}}_{\mathrm{n}}, \mathbf{Y}^{\mathrm{local}}_{\mathrm{n}} = \mathcal{F}_{\mathrm{crop}}(\mathbf{X}^{\mathrm{c}}_{\mathrm{n}}, \mathbf{Y}^{\mathrm{c}}_{\mathrm{n}}), \tag{10}$$

where $\mathcal{F}_{\mathrm{crop}}$ represents the local cropping operation.

We then construct a new training input by mixing the cropped novel sample $\mathbf{X}^{\mathrm{local}}_{\mathrm{n}}$ with the current base input

$\mathbf{X}_\mathrm{b}$. Unlike previous mixup methods [39, 64], which discard scene context, we argue that context information is essential for models to better recognize challenging novel objects and propose to preserve it. To achieve this, we extract the four corners in the XY plane for both $\mathbf{X}_\mathrm{n}^\mathrm{local}$ and $\mathbf{X}_\mathrm{b}$, and then select a pair of opposite corners between them:

$$\mathbf{L}_\mathrm{b}, \mathbf{L}_\mathrm{n}^\mathrm{local} = \mathcal{F}_\mathrm{pair}(\mathcal{F}_\mathrm{corner}(\mathbf{X}_\mathrm{b}), \mathcal{F}_\mathrm{corner}(\mathbf{X}_\mathrm{n}^\mathrm{local})), \quad (11)$$

where $\mathcal{F}_\mathrm{corner}$ extracts the four corner points in the XY projections, and $\mathcal{F}_\mathrm{pair}$ randomly selects an opposing corner pair. Possible pairs include, for example, the leftmost corner of $\mathbf{X}_\mathrm{b}$ with the rightmost corner of $\mathbf{X}_\mathrm{n}^\mathrm{local}$, or the uppermost corner of $\mathbf{X}_\mathrm{b}$ with the lowermost corner of $\mathbf{X}_\mathrm{n}^\mathrm{local}$. Here, $\mathbf{L}_\mathrm{b}$ and $\mathbf{L}_\mathrm{n}^\mathrm{local}$ are the coordinates of the selected corners. To align $\mathbf{X}_\mathrm{b}$ with $\mathbf{X}_\mathrm{n}^\mathrm{local}$ at the chosen corners, we compute a translation vector $\mathbf{T} = \mathbf{L}_\mathrm{b} - \mathbf{L}_\mathrm{n}^\mathrm{local}$ and apply it to translate $\mathbf{X}_\mathrm{n}^\mathrm{local}$, yielding the final mixed result. This ensures a close connection between the samples without losing context, which is crucial for effective novel class learning. Visualizations of the output can be found in Sec. 5.3, with further details in the supplementary material.

5. Experiments

5.1. Experimental Setup

Datasets. Our new evaluation benchmark builds on two datasets: 1) ScanNet200 [48] – This dataset extends the labeling scope of ScanNet [7] from 20 to 200 classes, adding finer-grained subclasses of existing categories and numerous new classes. 2) ScanNet++ [70] – Comprising 460 scenes with annotations for over 1000 unique classes, ScanNet++ is designed to capture a wide range of object types. To establish a comprehensive GFS-PCS benchmark, we selected the most frequent classes from each dataset, ensuring adequate representation across scenes. Our final benchmark includes 57 classes for ScanNet200 (with 40 novel classes) and 30 classes for ScanNet++ (with 18 novel classes). Full class lists are in the supplementary material. We follow the standard training/testing splits for each dataset, adhering to the preprocessing and augmentation settings from [63], where raw input points are voxelized at a 0.02m grid size. Notably, unlike prior GFS-PCS evaluations [66, 77], which test models on small blocks, we test on whole scenes to better simulate real-world scenarios.

Implementation Details. Our framework uses a straightforward segmentor with a backbone and a linear classification head, optimized for efficiency and simplicity. By default, we use Point Transformer V3 (PTv3) [63] as the backbone and 3D VLM RegionPLC [67]. The segmentor is first pre-trained on base classes of each dataset, after which we add a separate linear classification head for novel classes, enabling lightweight and efficient adaptation to novel classes in 20 fine-tuning epochs. The pre-training

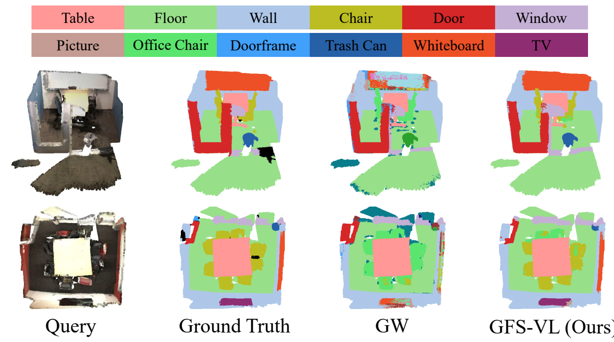


Figure 3. **Qualitative comparison between GW [66] and our GFS-VL on ScanNet200.** Class colors are shown at the top.

setting follows [63] with 800 epochs, and we use Adam optimizer for fine-tuning with learning rates: 0.001 for ScanNet200 and ScanNet, and 0.007 for ScanNet++. More details are in the supplementary material. For evaluation, we adopt metrics outlined in [66]: mean Intersection-over-Union (mIoU) for base classes (mIoU-B), novel classes (mIoU-N), all classes (mIoU-A), and the harmonic mean of mIoU-B and mIoU-N (HM) which captures overall performance while mitigating bias towards base classes [69].

5.2. Experimental Results

In our evaluation, we benchmark our method against baseline models attMPTI [77] and PIFS [4] following [66], along with the state-of-the-art GFS-PCS model GW [66] and FS-PCS model COSeg [1]. For fairness, all baseline models are retrained using the same backbone as our model. We also include a Fully Supervised model as an upper bound, obtained by fine-tuning an identical model to ours on ground-truth labels for both base and novel classes.

Following [56, 66], we evaluate two support scenarios, 5-shot and 1-shot, averaging performance across five randomly-seeded support set versions. As shown in Tab. 2, on ScanNet200, our model achieves substantial improvements across all metrics and support scenarios, including a 28.57% increase in HM and a 23.37% boost in mIoU-N over the closest baseline, GW, in the 5-shot setting. Qualitative comparisons in Fig. 3 further illustrate the superior segmentation accuracy of our model compared to GW. Similar trends are seen on ScanNet++ (see Tab. 3), where our model improves HM by 17.88% and mIoU-N by 12.79% compared to GW (1-shot). For the traditional ScanNet benchmark with only six novel classes, we report results in Tab. 4 using baseline performance from [66], where our model significantly surpasses all baselines with an impressive 34.94% gain in mIoU-N and 39.33% in HM for the 1-shot task.

The substantial and consistent gains across diverse datasets and metrics highlight our method's strong adaptability to diverse and complex novel classes by effectively integrating valuable few-shot samples with the semantic insights of 3D VLMs [9, 44]. In contrast, baseline models rely

Method	5-shot				1-shot			
	mIoU-B	mIoU-N	mIoU-A	HM	mIoU-B	mIoU-N	mIoU-A	HM
Fully Supervised	68.70	39.32	45.51	50.02	68.70	39.32	45.51	50.02
PIFS [4]	28.78	3.82	9.07	6.71	17.84	2.87	6.02	4.88
attMPTI [77]	37.13	4.99	11.76	8.79	54.84	3.28	14.14	6.17
COSeg [1]	57.67	5.21	16.25	9.54	47.03	4.03	13.09	7.42
GW [66]	59.28	8.30	19.03	14.55	55.23	6.47	16.74	11.56
GFS-VL (ours)	**67.57**	**31.67**	**39.23**	**43.12**	**68.48**	**29.18**	**37.45**	**40.92**

Table 2. **Comparisons of our method with baselines on the new ScanNet200 benchmark.** The best results are highlighted in **bold**.

Method	5-shot				1-shot			
	mIoU-B	mIoU-N	mIoU-A	HM	mIoU-B	mIoU-N	mIoU-A	HM
Fully Supervised	65.45	37.24	48.53	47.47	65.45	37.24	48.53	47.47
PIFS [4]	39.98	5.74	19.44	10.03	36.66	4.95	17.63	8.71
attMPTI [77]	55.89	4.19	24.87	7.78	53.16	3.55	23.40	6.66
COSeg [1]	59.34	6.96	27.91	12.45	58.49	6.24	27.14	11.26
GW [66]	51.35	11.03	27.16	18.15	46.71	6.63	22.66	11.59
GFS-VL (ours)	**60.05**	**21.66**	**37.02**	**31.82**	**61.39**	**19.42**	**36.21**	**29.47**

Table 3. **Comparisons of our method with baselines on the new ScanNet++ benchmark.** The best results are highlighted in **bold**.

Method	5-shot				1-shot			
	mIoU-B	mIoU-N	mIoU-A	HM	mIoU-B	mIoU-N	mIoU-A	HM
Fully Supervised	78.71	60.37	72.91	68.33	78.71	60.37	72.91	68.33
attMPTI [77]	16.31	3.12	12.35	5.21	12.97	1.62	9.57	2.88
PIFS [4]	35.14	3.21	25.56	5.88	35.80	2.54	25.82	4.75
CAPL [56]	38.22	14.39	31.07	20.88	38.70	10.59	30.27	16.53
GW [66]	40.18	18.58	33.70	25.39	40.06	14.78	32.47	21.55
GFS-VL (ours)	**78.30**	**51.22**	**69.75**	**61.91**	**78.56**	**49.72**	**69.45**	**60.88**

Table 4. **Comparisons of our method with baselines on the old ScanNet benchmark.** The best results are highlighted in **bold**.

solely on limited few-shot samples, constraining their capacity for novel class generalization and resulting in lower scores on the benchmarks. Moreover, the reduced performance on our new benchmarks emphasizes their value as a rigorous and comprehensive evaluation setting. These benchmarks challenge models to demonstrate robust generalization across diverse classes, fostering a deeper understanding of GFS-PCS for real-world applications.

5.3. Ablation Studies

We conducted ablation studies on the ScanNet200 [48] dataset using a single set of 5-shot support samples.

Effect of Design Components. We assessed the effectiveness of each module in GFS-VL in Tab. 5a. The baseline (1st row) shows performance using raw pseudo-labels directly from the 3D VLM. Adding Pseudo-label Selection (PS) in the 2nd row significantly improves pseudo-label quality, resulting in a clear performance boost. Introducing Adaptive Infilling (AI) further enhances results by effectively assigning labels to unlabeled regions of novel classes (3th row). Visualizations in Fig. 4 illustrate the quality improvements achieved with PS and AI, affirming their effectiveness. Lastly, Novel-Base Mix (NB-Mix), whether used with PS (4rd row) or combined with PS and AI (5th row), improves generalization by effectively integrating novel knowledge into the training samples.

Results with Different 3D VLMs. We tested GFS-VLwith two prominent 3D VLMs, RegionPLC [67] and Openscene [44], as shown in the 2nd and 4th rows of Tab. 5b. Compared to their zero-shot results (1st and 3rd rows), our approach achieves substantial gains by effectively leveraging few-shot samples to refine and enhance the noisy knowledge in each 3D VLM. This demonstrates our framework's flexibility and effectiveness across diverse 3D VLMs.

Impact of the Threshold in AI. Adaptive Infilling (AI) assigns novel class labels to unlabeled points based on similarity to adaptive prototypes, controlled by a threshold δ. In Tab. 5c, we explore different δ values and observe that $\delta = 0.9$ achieves the highest performance, which best balances enriching novel classes and maintaining label quality.

Effect of the Threshold in PS. Pseudo-label Selection (PS) refines pseudo-labels by retaining only the most reliable regions predicted by the 3D VLM, based on a threshold τ. In Tab. 5d, we evaluate the effect of varying τ. The method performs robustly across different τ values, with the optimal

PS	AI	NB-Mix	mIoU-B	mIoU-N	mIoU-A	HM
			65.50	22.30	31.40	33.28
✓			69.26	26.51	35.51	38.35
✓	✓		66.25	28.03	36.07	39.39
✓		✓	66.94	28.21	36.36	39.69
✓	✓	✓	67.42	31.81	39.30	43.22

(a)

3D VLM	mIoU-B	mIoU-N	mIoU-A	HM
RegionPLC [67]	46.97	23.77	28.65	31.56
Ours (RegionPLC)	67.42	31.81	39.30	43.22
Openscene [44]	53.07	15.16	23.14	23.58
Ours (Openscene)	68.56	20.09	30.29	31.07

(b)

δ	mIoU-B	mIoU-N	mIoU-A	HM
0.80	66.67	30.41	38.04	41.77
0.85	66.29	31.00	38.43	42.24
0.90	67.42	31.81	39.30	43.22
0.95	66.22	30.68	38.17	41.94

(c)

τ	mIoU-B	mIoU-N	mIoU-A	HM
0.5	67.19	31.33	38.88	42.73
0.6	67.42	31.81	39.30	43.22
0.7	66.69	30.56	38.17	41.92
0.8	68.06	30.67	38.54	42.28

(d)

n	mIoU-B	mIoU-N	mIoU-A	HM
1	67.40	27.13	35.61	38.68
2	67.95	27.71	36.18	39.36
3	66.94	28.21	36.36	39.69
4	67.84	27.80	36.23	39.44

(e)

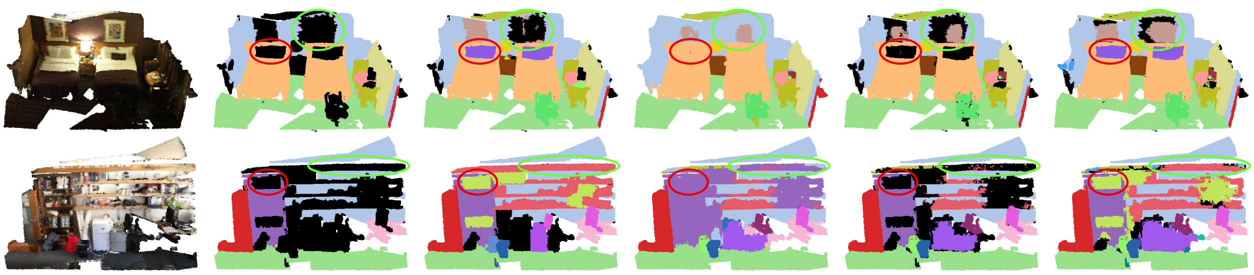
Backbone	mIoU-B	mIoU-N	mIoU-A	HM
PTv3 [63]	67.42	31.81	39.30	43.22
SCN [13]	61.85	31.94	38.24	42.13

(f)

Mix	mIoU-B	mIoU-N	mIoU-A	HM
Instance Mix	68.29	23.93	33.27	35.44
Mix3D [39]	68.50	24.80	34.00	36.42
NB-Mix	67.95	27.71	36.18	39.36

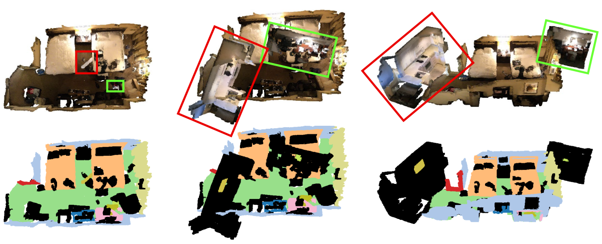
(g)

Table 5. **Ablation study.** (a) Effect of design components. (b) Results with different 3D VLMs. (c) Impact of the threshold in AI. (d) Effect of the threshold in PS. (e) Impact of different mix blocks. (f) Results with different backbones. (g) Comparison of mix strategies.



Query | Base Class Labels | Full Class labels | Raw Pseudo-Label | After PS | After AI

Figure 4. Visualization of the improvements in pseudo-label quality after applying Pseudo-label Selection (PS) and Adaptive Infilling (AI). Note that AI effectively discovers missed novel classes in the red circles and completes partial pseudo-labels in the green circles.



Instance Mix | Mix3D | Novel-Base Mix (Ours)

Figure 5. Visual illustration of mixing strategies. The red and green boxes represent the two novel samples mixed into the scene.

performance achieved at $\tau = 0.6$. This robustness indicates that PS effectively selects relevant reliable regions without being overly sensitive to the threshold choice.

Impact of Different Mix Blocks. In NB-Mix, we can adjust the number of novel blocks, denoted as n, used to augment each training sample. Tab. 5e shows stable performance (with the AI module disabled) across different block counts, suggesting that NB-Mix effectively integrates novel class information. By default, we use three blocks.

Results with Different Backbones. Besides varying 3D VLMs, we examined the effect of different backbones. We evaluated two backbones, PTv3 [63] and SparseConvNet (SCN) [13] in Tab. 5f. Our approach consistently performs well across both networks, confirming that GFS-VL is generalizable and not dependent on a specific backbone.

Comparison of Mix Strategies. We investigated alternative mix strategies for integrating novel support samples into training data in Tab. 5g. Specifically, we compared our NB-Mix with Instance Mix, which randomly inserts novel class objects from foreground masks into scenes, and Mix3D [39], which overlays two scenes for out-of-context augmentation. Fig. 5 shows the visual examples of these strategies. Our method outperforms these alternatives, highlighting the importance of preserving local context to effectively learn diverse and challenging novel classes.

6. Conclusion

This work introduces a GFS-PCS framework GFS-VL that synergizes dense but noisy pseudo-labels from 3D VLMs with accurate yet sparse few-shot samples, overcoming current GFS-PCS limitations in novel knowledge learning. GFS-VL utilizes prototype-guided pseudo-label selection to target high-quality regions and adaptive infilling to enrich pseudo-labels. Besides, the novel-base mix embeds few-shot samples into training scenes, preserving essential context for improved novel class learning. Identifying the limited diversity in current GFS-PCS evaluations, we introduce two benchmarks with broader, more diverse novel classes for more comprehensive generalization evaluation. GFS-VL achieves leading results and generalizes effectively across models and datasets, showing the potential of 3D VLMs in advancing GFS-PCS. We hope our method and benchmarks serve as a foundation for future research.

Acknowledgments

This work is supported by the Pioneer Centre for AI, DNRF grant number P1.

References

[1] Zhaochong An, Guolei Sun, Yun Liu, Fayao Liu, Zongwei Wu, Dan Wang, Luc Van Gool, and Serge Belongie. Rethinking few-shot 3D point cloud semantic segmentation. In *CVPR*, pages 3996–4006, 2024. 1, 2, 6, 7

[2] Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Min Wu, Ming-Ming Cheng, Ender Konukoglu, and Serge Belongie. Multimodality helps few-shot 3D point cloud semantic segmentation. In *ICLR*, 2025. 2

[3] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3D semantic parsing of large-scale indoor spaces. In *CVPR*, pages 1534–1543, 2016. 2, 3

[4] Fabio Cermelli, Massimiliano Mancini, Yongqin Xian, Zeynep Akata, and Barbara Caputo. Prototype-based incremental few-shot semantic segmentation. In *BMVC*, page 155, 2021. 3, 6, 7

[5] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. CLIP2Scene: Towards label-efficient 3D scene understanding by CLIP. In *CVPR*, pages 7020–7030, 2023. 2, 3

[6] Yunlu Chen, Vincent Tao Hu, Efstratios Gavves, Thomas Mensink, Pascal Mettes, Pengwan Yang, and Cees GM Snoek. PointMixup: Augmentation for point clouds. In *ECCV*, pages 330–345. Springer, 2020. 3

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 2, 3, 6

[8] Hao Deng, Kunlei Jing, Shengmei Chen, Cheng Liu, Jiawei Ru, Bo Jiang, and Lin Wang. Linnet: Linear network for efficient point cloud representation learning. In *NeurIPS*, 2024. 2

[9] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. PLA: Language-driven open-vocabulary 3D scene understanding. In *CVPR*, pages 7010–7019, 2023. 2, 3, 6

[10] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Lowis3D: Language-driven open-world instance-level 3D scene understanding. *IEEE TPAMI*, 46(12):8517–8533, 2024. 2, 3

[11] Jin Fang, Xinxin Zuo, Dingfu Zhou, Shengze Jin, Sen Wang, and Liangjun Zhang. LiDAR-Aug: A general rendering-based augmentation framework for 3D object detection. In *CVPR*, pages 4710–4720, 2021. 3

[12] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, pages 540–557, 2022. 2, 3

[13] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3D semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, pages 9224–9232, 2018. 8

[14] Sina Hajimiri, Malik Boudiaf, Ismail Ben Ayed, and Jose Dolz. A strong baseline for generalized few-shot semantic segmentation. In *CVPR*, pages 11269–11278, 2023. 3

[15] Jiawei Han, Kaiqi Liu, Wei Li, and Guangzhi Chen. Subspace prototype guidance for mitigating class imbalance in point cloud semantic segmentation. In *ECCV*, pages 255–272, 2024. 2

[16] Shuting He, Xudong Jiang, Wei Jiang, and Henghui Ding. Prototype adaption and projection for few-and zero-shot 3D point cloud semantic segmentation. *IEEE TIP*, 32:3199–3211, 2023. 2

[17] Mir Rayat Imtiaz Hossain, Mennatullah Siam, Leonid Sigal, and James J Little. Visual prompting for generalized few-shot segmentation: A multi-scale approach. In *CVPR*, pages 23470–23480, 2024. 2, 3

[18] 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. In *ICCV*, pages 22157–22167, 2023. 2, 3

[19] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, et al. ConceptFusion: Open-set multimodal 3D mapping. In *ICRA2023 Workshop on Pretraining for Robotics (PT4R)*, 2023. 3

[20] Jincen Jiang, Qianyu Zhou, Yuhang Li, Xinkui Zhao, Meili Wang, Lizhuang Ma, Jian Chang, Jian Jun Zhang, and Xuequan Lu. Pcotta: Continual test-time adaptation for multi-task point cloud understanding. In *NeurIPS*, 2024. 1

[21] Li Jiang, Shaoshuai Shi, and Bernt Schiele. Open-vocabulary 3D semantic segmentation with foundation models. In *CVPR*, pages 21284–21294, 2024. 2, 3

[22] Sihyeon Kim, Sanghyeok Lee, Dasol Hwang, Jaewon Lee, Seong Jae Hwang, and Hyunwoo J Kim. Point cloud augmentation with weighted local transformations. In *ICCV*, pages 548–557, 2021. 3

[23] Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. OneFormer3D: One transformer for unified point cloud segmentation. In *CVPR*, pages 20943–20953, 2024. 2

[24] 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 *CVPR*, pages 8500–8509, 2022. 2

[25] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *CVPR*, pages 8057–8067, 2022. 3

[26] Dogyoon Lee, Jaeha Lee, Junhyeop Lee, Hyeongmin Lee, Minhyeok Lee, Sungmin Woo, and Sangyoun Lee. Regularization strategy for point cloud via rigidly mixed sample. In *CVPR*, pages 15900–15909, 2021. 3

[27] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 2, 3

[28] Lei Li. Hierarchical edge aware learning for 3D point cloud. In *Computer Graphics International Conference*, pages 81–92. Springer, 2023. 2

[29] Ruihui Li, Xianzhi Li, Pheng-Ann Heng, and Chi-Wing Fu. Pointaugment: an auto-augmentation framework for point cloud classification. In *CVPR*, pages 6378–6387, 2020. 3

[30] Ruihuang Li, Zhengqiang Zhang, Chenhang He, Zhiyuan Ma, Vishal M Patel, and Lei Zhang. Dense multimodal alignment for open-vocabulary 3D scene understanding. In *ECCV*, 2024. 3

[31] Zhaoyang Li, Yuan Wang, Wangkai Li, Rui Sun, and Tianzhu Zhang. Localization and expansion: A decoupled framework for point cloud few-shot semantic segmentation. In *ECCV*, 2024. 2

[32] Chuandong Liu, Chenqiang Gao, Fangcen Liu, Pengcheng Li, Deyu Meng, and Xinbo Gao. Hierarchical supervision and shuffle data augmentation for 3D semi-supervised object detection. In *CVPR*, pages 23819–23828, 2023. 3

[33] Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. Learning orthogonal prototypes for generalized few-shot semantic segmentation. In *CVPR*, pages 11319–11328, 2023. 2, 3

[34] Weide Liu, Zhonghua Wu, Yang Zhao, Yuming Fang, Chuan-Sheng Foo, Jun Cheng, and Guosheng Lin. Harmonizing base and novel classes: A class-contrastive approach for generalized few-shot segmentation. *IJCV*, 132(4):1277–1291, 2024. 2, 3

[35] Zhe Liu, Jinghua Hou, Xinyu Wang, Xiaoqing Ye, Jingdong Wang, Hengshuang Zhao, and Xiang Bai. Lion: Linear group rnn for 3D object detection in point clouds. In *NeurIPS*, 2024. 1

[36] Zhihe Lu, Sen He, Da Li, Yi-Zhe Song, and Tao Xiang. Prediction calibration for generalized few-shot semantic segmentation. *IEEE TIP*, 32:3311–3323, 2023. 3

[37] Yongqiang Mao, Zonghao Guo, LU Xiaonan, Zhiqiang Yuan, and Haowen Guo. Bidirectional feature globalization for few-shot semantic segmentation of 3D point cloud scenes. In *3DV*, pages 505–514, 2022. 2

[38] Josh Myers-Dean, Yinan Zhao, Brian Price, Scott Cohen, and Danna Gurari. Generalized few-shot semantic segmentation: All you need is fine-tuning. *arXiv preprint*, 2021. 3

[39] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3D: Out-of-context data augmentation for 3D scenes. In *3DV*, pages 116–125. IEEE, 2021. 2, 3, 6, 8

[40] Dong Nie, Rui Lan, Ling Wang, and Xiaofeng Ren. Pyramid architecture for multi-scale processing in point cloud segmentation. In *CVPR*, pages 17284–17294, 2022. 2

[41] Zhenhua Ning, Zhuotao Tian, Guangming Lu, and Wenjie Pei. Boosting few-shot 3D point cloud segmentation via query-guided enhancement. In *ACM MM*, pages 1895–1904, 2023. 2

[42] Zhiyi Pan, Wei Gao, Shan Liu, and Ge Li. Distribution guidance network for weakly supervised point cloud semantic segmentation. In *NeurIPS*, 2024. 1

[43] Chunghyun Park, Yoonwoo Jeong, Minsu Cho, and Jaesik Park. Fast point transformer. In *CVPR*, pages 16949–16958, 2022. 2

[44] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. OpenScene: 3D scene understanding with open vocabularies. In *CVPR*, pages 815–824, 2023. 2, 3, 6, 7, 8

[45] Shoumeng Qiu, Jie Chen, Chenghang Lai, Hong Lu, Xiangyang Xue, and Jian Pu. Leveraging smooth deformation augmentation for LiDAR point cloud semantic segmentation. *IEEE Transactions on Intelligent Vehicles*, 9(2):3316–3329, 2024. 3

[46] 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 supervision. In *ICML*, pages 8748–8763, 2021. 3

[47] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, pages 7008–7024, 2017. 2, 3

[48] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3D semantic segmentation in the wild. In *ECCV*, pages 125–141, 2022. 2, 4, 6, 7

[49] Cristiano Saltori, Fabio Galasso, Giuseppe Fiameni, Nicu Sebe, Fabio Poiesi, and Elisa Ricci. Compositional semantic mix for domain adaptation in point cloud segmentation. *IEEE TPAMI*, 45(12):14234–14247, 2023. 2, 3

[50] Shivanand Venkanna Sheshappanavar, Vinit Veerendraveer Singh, and Chandra Kambhamettu. Patchaugment: Local neighborhood augmentation in point cloud classification. In *ICCV*, pages 2118–2127, 2021. 3

[51] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, pages 4077–4087, 2017. 2, 3

[52] Hongyu Sun, Qiuhong Ke, Yongcai Wang, Wang Chen, Kang Yang, Deying Li, and Jianfei Cai. Point-prc: A prompt learning based regulation framework for generalizable point cloud analysis. In *NeurIPS*, 2024. 1

[53] Yuliang Sun, Xudong Zhang, and Yongwei Miao. A review of point cloud segmentation for understanding 3D indoor scenes. *Visual Intelligence*, 2(1):14, 2024. 1

[54] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-vocabulary 3D instance segmentation. In *NeurIPS*, pages 68367–68390, 2023. 2, 3

[55] Anirud Thyagharajan, Benjamin Ummenhofer, Prashant Laddha, Om Ji Omer, and Sreenivas Subramoney. Segment-Fusion: Hierarchical context fusion for robust 3D semantic segmentation. In *CVPR*, pages 1236–1245, 2022. 5

[56] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *CVPR*, pages 11563–11572, 2022. 1, 2, 3, 6, 7

[57] Chih-Jung Tsai, Hwann-Tzong Chen, and Tyng-Luh Liu. Pseudo-embedding for generalized few-shot 3D segmentation. In *ECCV*, pages 383–400, 2024. 3

[58] Changshuo Wang, Meiqing Wu, Siew-Kei Lam, Xin Ning, Shangshu Yu, Ruiping Wang, Weijun Li, and Thambipillai Srikanthan. GPSFormer: A global perception and local structure fitting-based transformer for point cloud understanding. In *ECCV*, pages 75–92, 2024. 2

[59] Jiahui Wang, Haiyue Zhu, Haoren Guo, Abdullah Al Mamun, Cheng Xiang, and Tong Heng Lee. Few-shot point cloud semantic segmentation via contrastive self-supervision and multi-resolution attention. In *ICRA*, pages 2811–2817, 2023. 2

[60] 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. In *ICML*, pages 23318–23340, 2022. 2, 3

[61] Yizhou Wang, Longguang Wang, Qingyong Hu, Yan Liu, Ye Zhang, and Yulan Guo. Panoptic segmentation of 3D point clouds with gaussian mixture model in outdoor scenes. *Visual Intelligence*, 2(1):10, 2024. 1

[62] Lili Wei, Congyan Lang, Ziyi Chen, Tao Wang, Yidong Li, and Jun Liu. Generated and pseudo content guided prototype refinement for few-shot point cloud segmentation. In *NeurIPS*, pages 31103–31123, 2024. 2

[63] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer V3: Simpler faster stronger. In *CVPR*, pages 4840–4851, 2024. 2, 6, 8

[64] Yanhao Wu, Tong Zhang, Wei Ke, Congpei Qiu, Sabine Süsstrunk, and Mathieu Salzmann. Mitigating object dependencies: Improving point cloud self-supervised learning through object exchange. In *CVPR*, pages 23052–23061, 2024. 2, 3, 6

[65] Aoran Xiao, Jiaxing Huang, Dayan Guan, Kaiwen Cui, Shijian Lu, and Ling Shao. PolarMix: A general data augmentation technique for LiDAR point clouds. In *NeurIPS*, pages 11035–11048, 2022. 2, 3

[66] Yating Xu, Conghui Hu, Na Zhao, and Gim Hee Lee. Generalized few-shot point cloud segmentation via geometric words. In *ICCV*, pages 21506–21515, 2023. 1, 2, 3, 4, 6, 7

[67] Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. RegionPLC: Regional point-language contrastive learning for open-world 3D scene understanding. In *CVPR*, pages 19823–19832, 2024. 2, 3, 6, 7, 8

[68] 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 pretraining for open-world detection. In *NeurIPS*, pages 9125–9138, 2022. 3

[69] Han-Jia Ye, Hexiang Hu, and De-Chuan Zhan. Learning adaptive classifiers synthesis for generalized few-shot learning. *IJCV*, 129:1930–1953, 2021. 6

[70] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *ICCV*, pages 12–22, 2023. 2, 4, 6

[71] Jinglin Zhan, Tiejun Liu, Rengang Li, Jingwei Zhang, Zhaoxiang Zhang, and Yuntao Chen. Real-Aug: Realistic scene synthesis for LiDAR augmentation in 3D object detection. *arXiv preprint*, 2023. 2, 3

[72] Canyu Zhang, Zhenyao Wu, Xinyi Wu, Ziyu Zhao, and Song Wang. Few-shot 3D point cloud semantic segmentation via stratified class-specific attention based transformer network. In *AAAI*, pages 3410–3417, 2023. 2

[73] Guowen Zhang, Lue Fan, Chenhang He, Zhen Lei, Zhaoxiang Zhang, and Lei Zhang. Voxel mamba: Group-free state space models for point cloud based 3D object detection. In *NeurIPS*, 2024. 1

[74] Jinlai Zhang, Lyujie Chen, Bo Ouyang, Binbin Liu, Jihong Zhu, Yujin Chen, Yanmei Meng, and Danfeng Wu. PointCutMix: Regularization strategy for point cloud classification. *Neurocomputing*, 505:58–67, 2022. 3

[75] Junbo Zhang, Runpei Dong, and Kaisheng Ma. CLIP-FO3D: Learning free open-world 3D scene representations from 2D dense CLIP. In *ICCV*, pages 2048–2059, 2023. 2, 3

[76] Nan Zhang, Zhiyi Pan, Thomas H Li, Wei Gao, and Ge Li. Improving graph representation for point cloud segmentation via attentive filtering. In *CVPR*, pages 1244–1254, 2023. 2

[77] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Few-shot 3D point cloud semantic segmentation. In *CVPR*, pages 8873–8882, 2021. 1, 2, 6, 7

[78] Yu Zheng, Guangming Wang, Jiuming Liu, Marc Pollefeys, and Hesheng Wang. Spherical frustum sparse convolution network for lidar point cloud semantic segmentation. In *NeurIPS*, 2024. 2

[79] Yuchen Zhou, Jiayuan Gu, Tung Yen Chiang, Fanbo Xiang, and Hao Su. Point-sam: Promptable 3D segmentation model for point clouds. In *ICLR*, 2025. 3

[80] Guanyu Zhu, Yong Zhou, Rui Yao, and Hancheng Zhu. Cross-class bias rectification for point cloud few-shot segmentation. *IEEE TMM*, 25:9175–9188, 2023. 2

[81] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Jiaming Liu, Han Xiao, Chaoyou Fu, Hao Dong, and Peng Gao. No time to train: Empowering non-parametric networks for few-shot 3D scene segmentation. In *CVPR*, 2024. 2

[82] Ziyue Zhu, Qiang Meng, Xiao Wang, Ke Wang, Liujiang Yan, and Jian Yang. Curricular object manipulation in LiDAR-based object detection. In *CVPR*, pages 1125–1135, 2023. 2, 3