






Localization and Expansion: A Decoupled Framework for Point Cloud Few-shot Semantic Segmentation

Zhaoyang Li^{1*} , Yuan Wang^{1*} , Wangkai Li¹ , Rui Sun¹ , and Tianzhu Zhang^{1,2†} 

¹ MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China

² Deep Space Exploration Laboratory

{lizhaoyang, wy2016, lwklwk, issunrui}@mail.ustc.edu.cn, tz Zhang@ustc.edu.cn

Abstract. Point cloud few-shot semantic segmentation (PC-FSS) aims to segment targets of novel categories in a given query point cloud with only a few annotated support samples. The current top-performing prototypical learning methods employ prototypes originating from support samples to direct the classification of query points. However, the inherent fragility of point-level matching and the prevalent intra-class diversity pose great challenges to this cross-instance matching paradigm, leading to erroneous background activations or incomplete target excavation. In this work, we propose a simple yet effective framework in the spirit of Decoupled Localization and Expansion (DLE). The proposed DLE, including a structural localization module (SLM) and a self-expansion module (SEM), enjoys several merits. First, structural information is injected into the matching process through the agent-level correlation in SLM, and the confident target region can thus be precisely located. Second, more reliable intra-object similarity is harnessed in SEM to derive the complete target, and the conservative expansion strategy is introduced to reasonably constrain the expansion. Extensive experiments on two challenging benchmarks under different settings demonstrate that DLE outperforms previous state-of-the-art approaches by large margins.

Keywords: Point Cloud Analysis · Few-shot Learning · Semantic Segmentation

1 Introduction

As a fundamental computer vision task that boasts broad applications across autonomous driving, robotics, and beyond, point cloud semantic segmentation has propelled conspicuous achievements attributed to elaborate algorithms and well-established datasets. However, the capabilities of fully supervised segmentation models are restricted to predefined training categories, forming a close-set

[†] Corresponding author

^{*} Equal contribution

paradigm that severely limits their practical applications. In pursuit of human-like intelligence of recognizing novel categories from minimal examples, point cloud few-shot segmentation (PC-FSS) [36] has attracted increasing interest recently, which facilitates the segmentation of new category point clouds (query points) with only a few labeled point clouds (support points) without laborious model retraining.

Prototypical learning [22] has emerged as the mainstream paradigm of current top-performing PC-FSS methods [18, 35]. Specifically, these approaches condense the support features into a set of representative prototypes, which are then utilized to guide the classification of the query points. Though achieving promising progress, the significant scene discrepancies result in the direct matching between query points and support prototypes struggling to excavate the target, culminating in two primary types of deficiencies: **(1) Incorrect background activation.** The point-level matching primarily focuses on superficial and local consistencies, failing to integrate a broader consideration of semantic structural similarities. The cluttered background renders foreground prototypes susceptible to matching with background points that share local similarities as shown in Figure 1 (a). **(2) Incomplete foreground mining.** As illustrated in Figure 1 (b), pronounced intra-class diversities, such as the varieties of scale, pose, and so on, are common occurrences between the support and query objects. This cross-instance variation poses great challenges for support prototypes to furnish query target excavation with essential cues. To make matters worse, the mistake is inevitably amplified by inbuilt low-data regimes of PC-FSS, leading to sub-optimal results.

Some works [18, 33, 34] seek to mitigate this issue by equipping support prototypes with query information via the attention mechanism. Though the query-awareness of support prototypes can be improved in specific scenarios, the reliance on cross-instance prototypes and the fragile nature of point-level matching render these approaches highly sensitive to scene composition.

After an in-depth analysis of the above issues, in this paper, we propose decoupling the challenging PC-FSS into two intuitive procedures to coherently tackle the aforementioned different types of deficiencies, *i.e.*, **localization and expansion.** On the one hand, building upon the observation that pixels within the same object share higher similarity than those belonging to different objects [8, 13] (as shown in Figure 2 and Table 1), we deem that partially activated query object

features can serve as more reliable cues to infer complete target, as intra-object features are insusceptible to cross-instance variation. On the other hand, to alleviate the mismatches attributable to direct point-level comparison, we propose to compare the support prototypes and query points from the perspective of correlations on a set of representative reference features (referred to as *agents*). In this

Table 1: Quantitative measurement of intra- and inter-object similarity.

S3DIS		ScanNet	
Inter-object	Intra-object	Inter-object	Intra-object
0.25	0.41	0.14	0.39

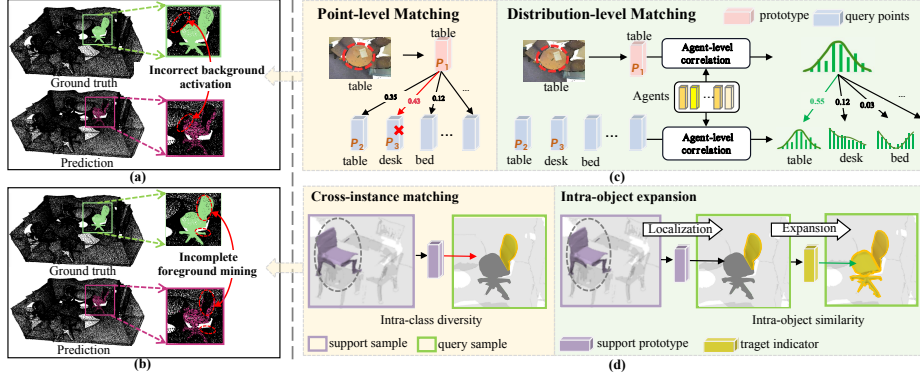


Fig. 1: Motivation of our method. (a), (b) Different types of segmentation deficiencies resulted from the inherent fragility of point-level matching and the prevalent intra-class diversity. (c) We employ distribution-level matching that incorporates structural information to replace point-level matching. (d) We leverage intra-object similarity to fully excavate targets, thereby circumventing the impact of intra-class diversity.

way, high-level semantic structural consistency is integrated to form a more comprehensive distribution-level matching paradigm, thereby achieving more precise target localization.

We propose a simple yet effective PC-FSS framework in the spirit of **Decoupled Localization and Expansion (DLE)**, which optimizes a structural localization module (SLM) and a self-expansion module (SEM). **To precisely locate the targets**, we introduce a set of semantic agents to rectify the direct point-level matching process. The main idea is, for each query point or support prototype, we can obtain the agent-level correlation (*i.e.*, a likelihood vector) by comparing it with all the agents. In essence, the agent-level correlation reflects the consensus among representative agents with a broader receptive field, thus implicitly integrating extensive structural information into the similarity measurement. Intuitively, the positive pair (*e.g.*, P_1 and P_2 in Figure 1 (c)) exhibits consistent similarity distribution across the agent sequence, while the prototype-point pairs (*e.g.*, P_1 and P_3 in Figure 1 (c)) that possess merely local similarities, lacking analogous agent-level correlation, are suppressed. Confident query target regions (albeit potentially incomplete) can thereby be located through the SLM. **To completely dig out the target region**, in SEM, the confident point features activated by SLM are collected as the target indicator via average pooling, which is then employed to capture the rest of the object areas based on the intra-object similarity. Although the target indicator is immune to the impact of intra-class diversity, it still risks erroneously expanding into the background due to potential similar interferences. To overcome this issue, we elegantly design a conservative expansion strategy to eliminate points that are likely to be background elements. Specifically, for each point within the region expanded by the target indicator, we identify its most similar counterpart within the entire query

point cloud as illustrated in Figure 1 (d). The points whose closest matches do not fall within the original region are filtered out, ensuring that the expansion remains within the consistent target areas.

The contributions of our method could be summarized as follows:

- We present a simple yet efficient PC-FSS framework in the spirit of Decoupled Localization and Expansion (DLE) to coherently tackle the different types of deficiencies existing in the current prototypical learning-based approaches.
- We introduce the Structural Localization Module (SLM) and Self-Expansion Module (SEM) for precise target localization and complete target excavation. They are elegantly integrated to alleviate the impact of scene discrepancies.
- Extensive experiments on two challenging benchmarks under different (*i.e.*, 1/2 way, 1/5 shot) settings demonstrate that our approach significantly outperforms previous state-of-the-art PC-FSS approaches.

2 Related Work

2.1 3D Point Cloud Semantic Segmentation

3D point cloud semantic segmentation is a fundamental computer vision task that aims to achieve point-wise classification of the given 3D point cloud on predefined categories. Benefiting from the advantages of deep learning [30, 32], remarkable progress has been achieved in parsing 3D scenes by recent deep-learning-based approaches [11, 14, 15, 17, 19, 24, 25, 28, 29]. PointNet [20] emerged as the pioneering end-to-end neural network designed for direct segmentation of raw point clouds, adeptly maintaining the permutation invariance characteristic of the input data. While PointNet offers a straightforward and efficient architecture, it fails to capture the critical local structures inherent in the spatial distribution of neighboring points. DGCNN [27] tackles this limitation by introducing the EdgeConv module, which effectively captures local structures. Despite their promising results, these methods fail to generalize to novel classes in the low-data regime. In this work, we utilize DGCNN as the backbone of our feature extractor to capture local structure features and propose our method to tackle the 3D point cloud semantic segmentation problem in a few-shot setting.

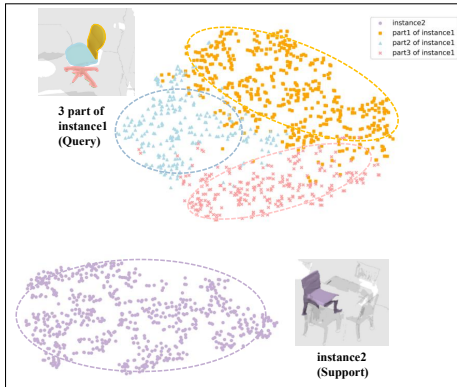


Fig. 2: T-SNE visualization highlights intra-class diversity and intra-object similarity.

2.2 Few-shot 3D Point Cloud Semantic Segmentation

Few-shot 3D Point Cloud Semantic Segmentation aims to perform point-wise classification on point clouds of previously unseen categories with only a handful of labeled point available. Given the limited availability of annotated examples, it is essential for Few-Shot Segmentation(FSS) tasks to fully leverage the categorical insights present in the support data to improve segmentation accuracy in the query set. Zhao et al. [36] pioneered the exploration of few-shot 3D point cloud semantic segmentation by introducing an attention-aware multi-prototype transductive inference approach, a novel strategy specifically tailored for classifying previously unseen categories with minimal annotated data. Originally conceived for few-shot image classification, ProtoNet [22] has been innovatively repurposed for 3D point cloud segmentation in the work of [36], which utilizes a DGCNN coupled with a linear projection module for feature extraction, producing support and query features that, after global average pooling to derive class prototypes, are matched using cosine similarity to generate predictions. To reduce contextual gaps between support prototypes and query features, [18] proposed PC-FSS method enhances segmentation by adapting support prototypes to the query context and holistically refining them guided by query features. Despite promising results, the significant intra-class diversities between support and query data persist, limiting the performance ceiling of 3D point cloud few-shot semantic segmentation. In this work, we handle the few-shot point cloud semantic segmentation by adopting the paradigm of [36]. We propose a simple yet effective PC-FSS framework in the spirit of **Decoupled Localization and Expansion (DLE)**, which leverages the internal similarity within query objects and utilizes holistic category-level cues from support data to effectively discern query objects.

2.3 Few-shot learning

Few-shot learning targets learning a new paradigm for a novel task with few samples. Previous research predominantly falls into two learning schemes: metric learning and meta-learning. Metric-learning based approaches [1, 7, 16, 22, 23, 26], employ a Siamese network [12] architecture to process pairs of support and query images, to learn a universal metric that assesses their similarity. Meta-learning based approaches [3, 4, 9, 10, 21, 24, 25, 31] facilitate swift model adaptation by meta-learning optimization algorithms using a limited number of samples. Based on the paradigm of existing few-shot learning methods, our research improves the performance of point cloud few-shot semantic segmentation with a novel and effective design.

3 Method

3.1 Problem Definition

The objective of few-shot point cloud semantic segmentation is to endow the segmentation model with the capacity for rapid generalization to novel categories

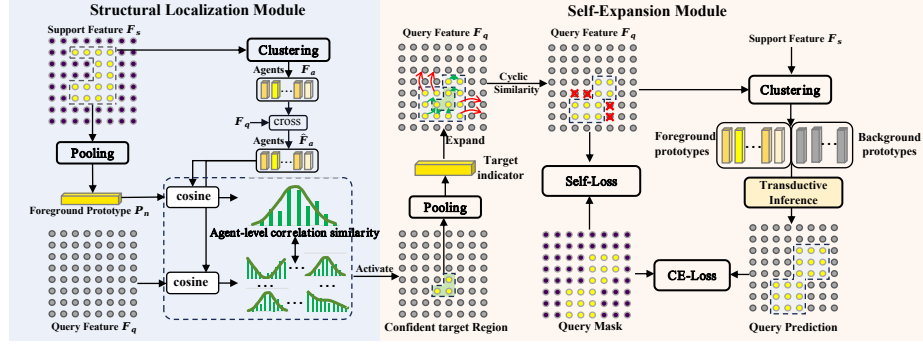


Fig. 3: Illustration of the proposed DLE. There are two main modules in DLE. The structural localization module is responsible for precisely locating confident target regions by introducing a set of semantically structure-aware agents. The self-expansion module expands the located target region to mine extensive query information further. These two modules collaboratively constitute a decoupled localization and expansion (DLE) framework for Point Cloud Few-Shot Segmentation.

with only a limited set of annotated samples. Following previous works, we adopt the widely used episodic paradigm [23]. Specifically, considering the training dataset $\mathcal{D}_{\text{train}}$ and the testing dataset $\mathcal{D}_{\text{test}}$, which possess non-overlapping target categories ($\mathcal{C}_{\text{train}} \cap \mathcal{C}_{\text{test}} = \emptyset$), we extract a series of subtasks from $\mathcal{D}_{\text{train}}$ to train the model, while evaluation is kept strictly separate, utilizing the testing dataset $\mathcal{D}_{\text{test}}$. Each subtask, also referred to as an episode, manifests as an N -way K -shot segmentation task within point cloud data. And each episode contains a support set denoted as $\mathcal{S} = \{(I_S^{n,k}, M_S^{n,k})\}$, where I is point cloud, M is mask, $k \in \{1, \dots, K\}$, $n \in \{1, \dots, N\}$, and a query set $Q = \{(I_Q, M_Q)\}$. During each episode, the model receives inputs comprising $\{I_S^{n,k}, M_S^{n,k}, I_Q\}$, where $I_S^{n,k}$ and $M_S^{n,k}$ are the point cloud and mask of the support set, respectively, and I_Q represents the query point cloud. The expected model output is the segmentation prediction for I_Q , with M_Q as the ground truth for validation. In the training phase, the model is optimized to infer the segmentation of the query set I_Q based on the support set S , guided by the ground truth mask M_Q . During testing, the generalization capabilities of the model are assessed using tasks derived from $\mathcal{D}_{\text{test}}$. To simplify the description, we describe our approach in a 1-way 1-shot setting.

3.2 Overview

As illustrated in Fig. 3, the proposed framework tailored for Decoupled Localization and Expansion PC-FSS comprises two procedures, i.e., (1) confident target region localization, (2) self-expansion, respectively handled by the structural localization module (SLM) and the self-expansion module (SEM). We locate the most confident target region of the novel class under the support guidance via distribution-level matching in procedure (1). Subsequently, in procedure (2),

the self-expansion module expands this initially partial region, enhancing it to more fully capture the target area. Through procedures (1) and (2), we obtain a pseudo query mask indicating the partial target region of the novel categories. The features of this region are then utilized to derive multiple prototypes, facilitating transductive inference [36], for producing the final segmentation results. The details are as follows.

3.3 Structural localization module

Considering the significant intra-class diversity between support and query, the direct point-level matching between support prototypes and query points in the procedure (1) for locating the query target can be challenging and susceptible to incorrect matches. To address this issue, we design a structural localization module that innovatively shifts the paradigm from direct point-level matching to a more stringent distribution-level matching through a set of semantically aware agents. Intuitively, agents serve as a pivotal bridge in measuring the similarity between support and query, necessitating a clear perception of semantic information. Therefore, the semantically-aware agents provide a crucial foundation for precise matching in subsequent stages. To endow agents with semantic awareness, we obtain a set of agents \mathbf{F}_a by applying farthest point sampling and K-means clustering on support foreground features, ensuring a diverse and effective setup. Intuitively, the farthest points in feature space can inherently encapsulate diverse aspects of one class (i.e., capturing fine-grained attributes). Given the initial agents features $\mathbf{F}_a = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{N_a}]$ ($\mathbf{a}_i \in \mathbb{R}^{1 \times d}$ indicates the i -th agent feature), we further engage these agents in a cross-attention mechanism with the query features $\mathbf{F}_q \in \mathbb{R}^{M \times d}$ (M signifies the point count and d the number of the feature channels) to adapt them to the query context. In the cross-attention module, the *queries* (\mathbf{Q}) is derived from the agents, while the *keys* (\mathbf{K}) and *values* (\mathbf{V}) are sourced from the query features \mathbf{F}_q , formally:

$$\mathbf{Q}_i = \mathbf{a}_i \mathbf{W}^{\mathcal{Q}}, \quad \mathbf{K}_j = \mathbf{f}_j \mathbf{W}^{\mathcal{K}}, \quad \mathbf{V}_j = \mathbf{f}_j \mathbf{W}^{\mathcal{V}}, \quad (1)$$

among which, $\mathbf{W}^{\mathcal{Q}} \in \mathbb{R}^{d \times d_k}$, $\mathbf{W}^{\mathcal{K}} \in \mathbb{R}^{d \times d_k}$, $\mathbf{W}^{\mathcal{V}} \in \mathbb{R}^{d \times d_v}$ are linear projections. For the i -th \mathbf{Q}_i , we calculate the attention weights via dot-product between \mathbf{Q}_i and all other *keys*:

$$w_{i,j} = \frac{\exp(\beta_{i,j})}{\sum_{j=1}^M \exp(\beta_{i,j})}, \quad \beta_{i,j} = \frac{\mathbf{Q}_i \mathbf{K}_j^T}{\sqrt{d_k}}, \quad (2)$$

where $\sqrt{d_k}$ is a scaling factor. The context-aware agent features are obtained via the weighted sum over all values:

$$\hat{\mathbf{a}}_i = \sum_{j=1}^M w_{i,j} \mathbf{V}_j, \quad (3)$$

Then a feed-forward network (FFN) is applied to obtain the semantic-aware agents $\hat{\mathbf{F}}_a = [\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, \dots, \hat{\mathbf{a}}_{N_a}]$. For a given support feature $\mathbf{F}_s \in \mathbb{R}^{M \times d}$ and the

corresponding mask $\mathbf{M}_s \in \mathbb{R}^{M \times 1}$, the prototype for category n is derived through masked average pooling (MAP) as follows:

$$\mathbf{P}_n = \text{MAP}(\mathbf{F}_s, \mathbf{M}_s). \quad (4)$$

Given the semantic-aware agents $\hat{\mathbf{F}}_a \in \mathbb{R}^{N_a \times d}$ and the category prototype $\mathbf{P}_n \in \mathbb{R}^{1 \times d}$, we can define the relation matrix \mathbf{W}_1 for the category prototype \mathbf{P}_n , which encodes the agent-level correlation, formally:

$$\mathbf{W}_1 = \text{Sim}(\mathbf{P}_n, \hat{\mathbf{F}}_a), \quad (5)$$

the *Sim* denotes the cosine similarity. Similarly, we also define the relation matrix \mathbf{W}_2 for the query features \mathbf{F}_q , which encodes the agent-level correlation, formally:

$$\mathbf{W}_2 = \text{Sim}(\mathbf{F}_q, \hat{\mathbf{F}}_a). \quad (6)$$

Given the correlation-encoded matrices $\mathbf{W}_1 \in \mathbb{R}^{1 \times N_a}$ and $\mathbf{W}_2 \in \mathbb{R}^{M \times N_a}$, a fully connected (FC) layer is initially employed to enhance and integrate features. Then the correlation matrix $\mathbf{C} \in \mathbb{R}^{1 \times M}$ is calculated as follows:

$$\mathbf{C} = \text{Sim}(FC(\mathbf{W}_1), FC(\mathbf{W}_2)), \quad (7)$$

$$\mathbf{M}_{q\tau} = \mathbb{1}^\tau(\mathbf{C}), \quad \mathbb{1}^\tau(x) = \begin{cases} 1, & x > \tau \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

In N-way settings ($N > 1$), we obtain N groups of agents $\{\mathbf{F}_a^i\}_{i=1,2,\dots,N}$ ($\mathbf{F}_a^i \in \mathbb{R}^{N_a \times d}$) from support foreground features, which are then concatenated to form a comprehensive set of agents $\mathbf{F}_A \in \mathbb{R}^{(N \times N_a) \times d}$. We then compute the correlation matrix \mathbf{C} of class prototype \mathbf{P}_n and query features \mathbf{F}_q based on their similarity distributions on \mathbf{F}_A . Subsequently, we employ a comparatively high threshold, denoted as τ , to predominantly identify class-relevant regions while effectively suppressing activation in background areas. In our experimental setup, τ is set to 0.7. If a point exceeds the threshold for multiple categories, it is assigned to the one with the highest similarity.

3.4 Self-expansion module

To further mine extensive query information from the confident target regions $\mathbf{M}_{q\tau}$ generated by Eq. (8), we introduce the self-expansion module (SEM), specifically designed to expand the coverage from localized confident regions to broader target areas by leveraging query intra-object similarity. In specific, utilizing the query features \mathbf{F}_q and the confident activation mask $\mathbf{M}_{q\tau}$, we obtain the target prototype \mathbf{P}_t by applying (mask average pooling) MAP to query feature \mathbf{F}_q , formally:

$$\mathbf{P}_t = \text{MAP}(\mathbf{F}_q, \mathbf{M}_{q\tau}). \quad (9)$$

Given $\mathbf{F}_q \in \mathbb{R}^{M \times d}$ and $\mathbf{P}_t \in \mathbb{R}^{1 \times d}$, we can calculate their similarity matrix as follows:

$$\mathbf{W}_3 = \text{Sim}(\mathbf{P}_t, \mathbf{F}_q), \quad (10)$$

where $\mathbf{W}_3 \in \mathbb{R}^{1 \times M}$ indicates every query point's similarity with \mathbf{P}_t . Then we can expand the confident region based on the matrix \mathbf{W}_3 :

$$\mathbf{M}_{e\theta} = \mathbb{1}^\theta(\mathbf{W}_3), \quad \mathbb{1}^\theta(x) = \begin{cases} 1, & x > \theta \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

In our experimental setup, θ is set to 0.8. If a point exceeds the threshold for multiple categories, it is assigned to the one with the highest similarity.

Although the intra-class similarity within the query is reliable, false matches are still inevitable during the target area expansion process. We adopt a **conservative expansion strategy**, employing a bidirectional confirmation of similarity to reduce false matches further. More precisely, for each point p in $\mathbf{M}_{e\theta}$ that is newly expanded relative to $\mathbf{M}_{q\tau}$, we conduct a cyclic similarity check. This entails identifying the p' in the query that exhibits the highest similarity to p . If p' is located within the original confident region $\mathbf{M}_{q\tau}$, it indicates a cyclical consistency between p and points in $\mathbf{M}_{q\tau}$, thus retaining p as part of the expanded region. Conversely, if p' falls outside $\mathbf{M}_{q\tau}$, p is likely a misactivated background point and is consequently discarded by setting its corresponding value in $\mathbf{M}_{e\theta}$ to 0. This process ensures that every newly expanded point in the query maintains cyclic consistency with the initially identified confident activation points. This verification helps prevent the target area's expansion from incorrectly extending into background points, thereby securing a relatively pure target region.

Owing to the inherent clutter in the background, we refrained from applying a similar procedure directly to identify background regions in the query. Here, we obtain L background prototypes $\mathbf{P}_b \in \mathbb{R}^{L \times d}$ by employing K-means on support background features, as multiple prototypes enhance the representation of complex and cluttered backgrounds. But directly utilizing the background prototype \mathbf{P}_b to guide the segmentation of query backgrounds can lead to sub-optimal results due to contextual gaps between the support and query. Consequently, we use the mask cross attention (MCA) mechanism that leverages background features within the query to guide \mathbf{P}_b in adapting to the query's background. In summary, for background prototype \mathbf{P}_b , query features \mathbf{F}_q , and the obtained target mask $\mathbf{M}_{e\theta}$, we perform a mask cross attention operation, akin to Mask2Former [5], where *queries* (\mathbf{Q}) are derived from \mathbf{P}_b , *keys* (\mathbf{K}) and *values* (\mathbf{V}) are sourced from the query features \mathbf{F}_q , $\mathbf{M}_{e\theta}$ is utilized to execute a masking operation, please refer to the Supplementary Material for more details. Through the MCA operation, we confine the attention to the query background regions, enabling P_b to adapt more effectively to the background of the query while mitigating interference from foreground points. To ensure precision in foreground predictions and penalize mispredictions of background points, we devise a self-loss as follows:

$$Loss_{self} = \frac{\sum_i (\mathbf{M}_{e\theta}(i) = 1 \wedge M_q(i) = 0)}{\sum_i (M_q(i) = 0)}. \quad (12)$$

where M_q is the query mask, and i indicates coordinate positions. Finally, we follow AttMPTI [36], integrating a multi-prototype transductive inference approach to derive the ultimate segmentation results.

Table 2: Results on **S3DIS** dataset using mean-IoU metric (%). S^i denotes the split i is used for testing. The best results are shown in **bold**

Method	1-way						2-way					
	1-shot			5-shot			1-shot			5-shot		
	S^0	S^1	Mean	S^0	S^1	Mean	S^0	S^1	Mean	S^0	S^1	Mean
ProtoNet _[NeurIPS'17] [22]	66.18	67.05	66.62	70.63	72.46	71.55	48.39	49.98	49.19	57.34	63.22	60.28
MPTI _[CVPR'21] [36]	64.13	65.33	64.73	68.68	68.04	68.45	52.27	51.58	51.88	58.93	60.65	59.75
AttMPTI _[CVPR'21] [36]	66.27	69.41	67.84	78.62	80.74	79.68	53.77	55.94	54.96	61.67	67.02	64.35
SCAT _[AAAI'23] [33]	69.37	70.56	69.96	70.13	71.36	70.74	54.92	56.74	55.83	64.24	69.03	66.63
ProtoNet+QGE _[MM'23] [18]	69.39	72.33	70.84	74.07	75.34	74.71	48.98	52.62	50.8	58.85	64.26	61.56
AttMPTI+QGE _[MM'23] [18]	74.30	77.62	75.96	81.86	82.39	82.13	58.85	60.29	59.57	66.56	79.46	69.01
Ours	76.54	78.8	77.67	83.15	83.23	83.19	61.34	63.58	62.46	67.92	74.79	71.36

Table 3: Results on **ScanNet** dataset using mean-IoU metric (%). S^i denotes the split i is used for testing. The best results are shown in **bold**

Method	1-way						2-way					
	1-shot			5-shot			1-shot			5-shot		
	S^0	S^1	Mean	S^0	S^1	Mean	S^0	S^1	Mean	S^0	S^1	Mean
ProtoNet _[NeurIPS'17] [22]	55.98	57.81	56.90	59.51	63.46	61.49	30.95	33.92	32.44	42.01	45.34	43.68
MPTI _[CVPR'21] [36]	52.13	57.59	54.86	62.13	63.73	62.93	36.14	39.27	37.71	43.59	46.90	45.25
AttMPTI _[CVPR'21] [36]	56.67	59.79	58.23	66.70	70.29	68.50	40.83	42.55	41.69	50.32	54.00	52.16
SCAT _[AAAI'23] [33]	56.49	59.22	57.85	65.19	66.82	66.00	45.24	45.90	45.57	55.38	57.11	56.24
ProtoNet+QGE _[MM'23] [18]	57.40	59.31	58.36	60.83	66.01	63.42	37.18	39.28	38.23	44.11	47.01	45.56
AttMPTI+QGE _[MM'23] [18]	59.06	61.66	60.36	66.88	72.17	69.53	43.10	46.79	44.95	51.91	57.21	54.56
Ours	61.58	63.08	62.33	68.18	72.72	70.45	46.12	49.82	47.97	54.34	59.79	57.07

4 Experiments

4.1 Datasets and Evaluation Metric

Datasets. Our experiments are conducted on two benchmark datasets for 3D point clouds: (1) **S3DIS** [2] which comprises 272 point clouds of diversely styled indoor environments including lobbies, hallways, offices, and pantries. This dataset is annotated with 12 semantic classes and an additional background class for clutter. (2) **ScanNet** [6] which consists of 1,513 point clouds derived from 707 distinct indoor scenes. Annotations within this dataset categorize 20 semantic classes, with an additional class assigned for regions without annotation. Within each dataset, classes are divided into two distinct subsets for cross-validation purposes, specifically S^0 and S^1 . These subsets are alternately used as the test classes $\mathcal{C}_{\text{test}}$ and the training classes $\mathcal{C}_{\text{train}}$. Further details regarding the partitioning process are provided in the appendix. Consistent with the preprocessing strategy of previous work [18], we partition rooms into 1m x 1m blocks, from which 2048 points are randomly sampled for each block.

Evaluation Metric. Following previous works [18, 36], Mean Intersection-over-Union (mean-IoU) is adopted as our evaluation metrics for all experiments. The Intersection over Union (IoU) for individual classes is computed as $\frac{TP}{TP+FN+FP}$, wherein TP , FN , and FP represent the counts of true positives, false negatives, and false positives, respectively. In the context of few-shot learning, the mean-

IoU is derived by averaging the IoU across all testing classes in the unseen category set $\mathcal{C}_{\text{test}}$.

4.2 Implementation Details

Our approach is implemented using the PyTorch framework, and all experiments are conducted on one NVIDIA GeForce RTX 3090 GPU. The feature extraction module is pre-trained on the training set $\mathcal{D}_{\text{train}}$ for 100 epochs with a batch size of 32, employing the Adam optimizer with a learning rate of 0.001. After pre-training, we train our model using the Adam optimizer, setting the initial learning rate to 0.0001 for the feature extractor—initialized with pre-trained weights—and 0.001 for the other modules within the architecture, both of them are halved every 5000 iterations. Following [36], models undergo 40,000 iterations throughout the training process. During each iteration, we randomly sample an episode, applying Gaussian noise and arbitrary z-axis rotations to augment the point clouds in both the support and query sets.

4.3 Comparison with State-of-the-Art Methods

S3DIS. In Table 2, we compare our proposed DLE with the state-of-the-art few-shot point cloud semantic segmentation methods. It can be observed that our DLE outperforms all previous models under all settings. Specifically, our approach achieves 77.67% and 62.46% mIoU in the 1-way 1-shot and 2-way 1-shot settings, surpassing the most competitive QGE [18] by 1.71% and 2.89%, respectively. Furthermore, the improvements are more pronounced in the 1-shot setting than in the 5-shot due to the scarcity of support information, which amplifies intra-class diversity issue. To address this, our method leverages distribution-level matching and better intra-object similarity within the query, thereby achieving more precise segmentation.

ScanNet. In Table 3, we compare results on the more challenging ScanNet dataset, which features diverse room types. Our DLE method still outperforms the best existing method, with gains of 1.97% and 3.02% mIoU in the 1-way 1-shot and 2-way 1-shot settings, respectively. Our model shows greater improvements on ScanNet due to its significant intra-class diversity, as shown in Table 1. This aligns with the issues our model addresses, highlighting its robustness in complex environments. Additionally, Fig. 4 shows prediction visualizations, demonstrating the visual superiority of our method.

4.4 Ablation Study

A comprehensive series of ablation studies were conducted on the S3DIS S^0 split under the 1-way 1-shot setting, employing AttMPTI [36] as our baseline model, to analyze each component of the proposed DLE.

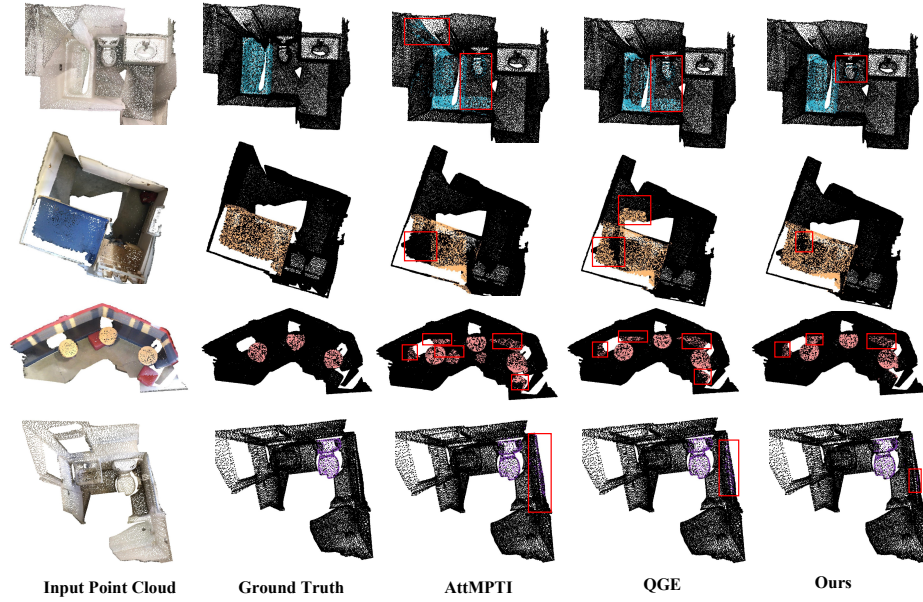


Fig. 4: Qualitative results of our method in 1-way 1-shot point cloud few-shot segmentation on ScanNet dataset in comparison to the ground truth, AttMPTI and QGE. The target classes from top to bottom are "bathhtub" (first row), "bed" (second row), "table" (third row) and "toilet" (last row).

Effectiveness of the structural localization module.

The comparison between the 1st and 5th rows of Table 4a, we can observe a significant performance lift, i.e., 7.99% in mIoU, demonstrating that our proposed semantically structure-aware agents effectively address the issue of mismatch, thereby accurately identifying discriminative foreground regions in the query point cloud. This provides a solid foundation for subsequent utilization of reliable intra-class similarity within the query to guide its segmentation.

Analysis of Agent Generation. To evaluate the effectiveness of different agent generation methods, we compare several intuitive initialization approaches in Table 4b. The *Random* method selects N_a point features randomly from the support features. *Learnable* initializes agents randomly but makes them trainable. *Cluster(all)* and *Cluster(fore)* cluster N_a agents from all support points

Table 4: Ablation study results. Experiments are conducted on S3DIS- S^0 for 1-way 1-shot setting.

SLM	SEM		mIoU	Δ
	Self-loss	Cycle		
			66.27	+0.0
	✓		67.9	+1.63
		✓	70.61	+4.34
	✓	✓	71.56	+5.29
✓			74.26	+7.99
✓	✓	✓	76.54	+10.27

(a) Ablation studies of the proposed SLM and SEM modules.

Init	mIoU
Random	72.1
Learnable	73.08
Cluter(all)	75.56
Cluster(fore)	76.54

(b) Ablation studies on different approaches for agents initialization.

and foreground point features, respectively. The Cluster-based methods consistently outperform others due to their ability to identify meaningful patterns and structures in the data. Specifically, clustering within foreground points is more effective, avoiding background noise and yielding agents that better reflect the structured semantic information of the foreground. This provides a solid foundation for using agents to establish distribution-level matching between support and query.

Investigation of the agent-based distribution-level matching.

To more intuitively demonstrate the effectiveness of our proposed agent-based distribution level matching, we conduct T-SNE visualizations both in the original feature space (Fig. 5 (a)) and in an agent-dimension-modified space (Fig. 5 (b)) for support prototypes and query features. In the agent-dimension-modified space, each point’s features reveal its similarity to our semantically aware agents, allowing us to more accurately separate foreground from background with finer detail. This approach not only clarifies the distinction between foreground and background but also aligns the support foreground prototype more closely with the query foreground, enabling a more accurate selection of query foreground.

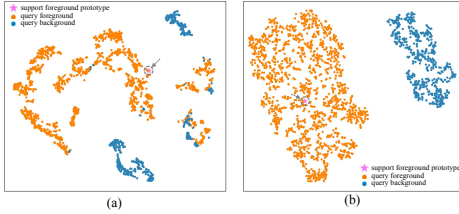


Fig. 5: T-SNE visualization of the feature distribution of the support foreground prototype, query foreground and query background in the original feature space (a) and in the agent-dimension-modified space (b).

Effectiveness of the self-expansion module. As shown in Table 4a, the full SEM module improves performance by 5.29% (comparing the 1st and 4th rows). Moreover, the addition of the self-mining module boosts performance by 2.28% mIoU (comparing the 5th and 6th rows) based on the structured disambiguation localization module. This enhancement is due to SEM’s bidirectional confirmation of similarity check and a tailored loss function, which leverages intra-object similarity within the query to excavate more target regions accurately.

Decoupled Localization and Expansion Paradigm Discussion. Since FSS models are trained on limited data, the feature space misaligns for new classes, causing poor feature representation and scattered features (intra-class diversities). This hinders direct matching between query points and support prototypes, leading to 1) incorrect background activation and 2) incomplete foreground mining. Figure 6 illustrates our solutions: (a) shows that direct matching causes false matches, reducing coverage and precision rates for foreground points;

Table 5: Ablation studies of the number of agents.

N_a	50	100	150	200
mIoU	75.79	76.54	76.19	75.06

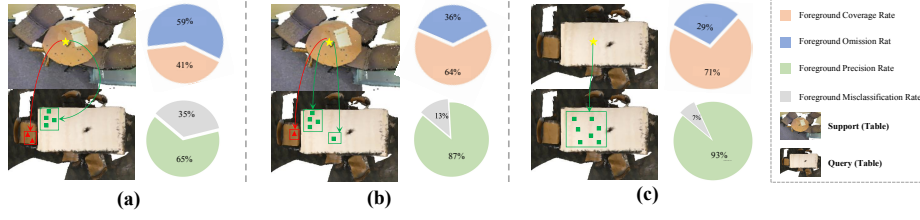


Fig. 6: Illustration of point correspondence and statistics data visualization. The red arrows indicate the false matches and the green arrows indicate the correct matches. **Foreground Coverage Rate** is defined as the proportion of selected foreground points to all foreground points. **Foreground Precision Rate** is defined as the proportion of correctly identified foreground points among all selected points.

(b) demonstrates our distribution-level matching approach, which improves coverage and precision by reducing false matches; (c) employs query region prototypes aggregation for better intra-object similarity, achieving higher foreground coverage and matching precision.

Hyperparameter Evaluations. Quantitative experiments are conducted to find a suitable number of agents N_a and thresholds τ and θ . As shown in Table 5, increasing the number of agents improves performance by enabling finer and more comprehensive matching. However, too many agents introduce unnecessary and fragmented disturbances and reduce performance, with 100 agents found to be optimal. Fig. 7 illustrates the model’s performance (mIoU) on the S3DIS dataset for different threshold settings, with optimal performance at $\tau = 0.7$ and $\theta = 0.8$, respectively.

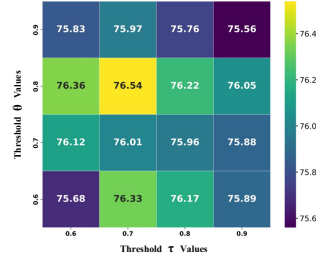


Fig. 7: Ablation results of thresholds τ and θ .

5 Conclusion

In this paper, we propose a novel Decoupled Localization and Expansion (DLE) approach to solve the **Incorrect background activation** and **Incomplete foreground mining** challenges in PC-FSS. Extensive experimental results on two standard benchmarks demonstrate that DLE performs favorably against state-of-the-art PC-FSS methods.

Acknowledgements

This work was partially supported by the National Defense Basic Scientific Research Program (JCKY2021130B016), National Nature Science Foundation of China (Grant 62021001), and Youth Innovation Promotion Association CAS.

References

1. Allen, K., Shelhamer, E., Shin, H., Tenenbaum, J.: Infinite mixture prototypes for few-shot learning. In: International conference on machine learning. pp. 232–241. PMLR (2019)
2. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1534–1543 (2016)
3. Bertinetto, L., Henriques, J.F., Torr, P.H., Vedaldi, A.: Meta-learning with differentiable closed-form solvers. arXiv preprint arXiv:1805.08136 (2018)
4. Cai, Q., Pan, Y., Yao, T., Yan, C., Mei, T.: Memory matching networks for one-shot image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4080–4088 (2018)
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
6. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
7. Doersch, C., Gupta, A., Zisserman, A.: Crosstransformers: spatially-aware few-shot transfer. *Advances in Neural Information Processing Systems* **33**, 21981–21993 (2020)
8. Fan, Q., Pei, W., Tai, Y.W., Tang, C.K.: Self-support few-shot semantic segmentation. In: European Conference on Computer Vision. pp. 701–719. Springer (2022)
9. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference on machine learning. pp. 1126–1135. PMLR (2017)
10. Gordon, J., Bronskill, J., Bauer, M., Nowozin, S., Turner, R.E.: Meta-learning probabilistic inference for prediction. arXiv preprint arXiv:1805.09921 (2018)
11. Huang, Q., Wang, W., Neumann, U.: Recurrent slice networks for 3d segmentation of point clouds. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2626–2635 (2018)
12. Koch, G., Zemel, R., Salakhutdinov, R., et al.: Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop. vol. 2. Lille (2015)
13. Koffka, K.: Principles of Gestalt psychology, vol. 44. Routledge (2013)
14. Lai, X., Liu, J., Jiang, L., Wang, L., Zhao, H., Liu, S., Qi, X., Jia, J.: Stratified transformer for 3d point cloud segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8500–8509 (2022)
15. Landrieu, L., Simonovsky, M.: Large-scale point cloud semantic segmentation with superpoint graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4558–4567 (2018)
16. Li, H., Eigen, D., Dodge, S., Zeiler, M., Wang, X.: Finding task-relevant features for few-shot learning by category traversal. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1–10 (2019)
17. Li, Z., Li, W., Mai, H., Zhang, T., Xiong, Z.: Enhancing cell detection in histopathology images: A vit-based u-net approach. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 150–160. Springer (2023)

18. Ning, Z., Tian, Z., Lu, G., Pei, W.: Boosting few-shot 3d point cloud segmentation via query-guided enhancement. In: *Proceedings of the 31st ACM International Conference on Multimedia*. pp. 1895–1904 (2023)
19. Pan, Y., Luo, N., Sun, R., Meng, M., Zhang, T., Xiong, Z., Zhang, Y.: Adaptive template transformer for mitochondria segmentation in electron microscopy images. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 21474–21484 (2023)
20. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 652–660 (2017)
21. Rusu, A.A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., Hadsell, R.: Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960* (2018)
22. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. *Advances in neural information processing systems* **30** (2017)
23. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. *Advances in neural information processing systems* **29** (2016)
24. Wang, Y., Luo, N., Zhang, T.: Focus on query: Adversarial mining transformer for few-shot segmentation. *arXiv preprint arXiv:2311.17626* (2023)
25. Wang, Y., Sun, R., Zhang, T.: Rethinking the correlation in few-shot segmentation: A buoys view. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7183–7192 (2023)
26. Wang, Y., Sun, R., Zhang, Z., Zhang, T.: Adaptive agent transformer for few-shot segmentation. In: *European Conference on Computer Vision*. pp. 36–52. Springer (2022)
27. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)* **38**(5), 1–12 (2019)
28. Wang, Z., Rao, Y., Yu, X., Zhou, J., Lu, J.: Semaffinet: Semantic-affine transformation for point cloud segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11819–11829 (2022)
29. Wangkai, L., Zhaoyang, L., Rui, S., Huayu, M., Naisong, L., Wang, Y., Yuwen, P., Guoxin, X., Huakai, L., Zhiwei, X., et al.: Maunet: Modality-aware anti-ambiguity u-net for multi-modality cell segmentation. In: *Competitions in Neural Information Processing Systems*. pp. 1–12. PMLR (2023)
30. Wu, W., Qi, Z., Fuxin, L.: Pointconv: Deep convolutional networks on 3d point clouds. In: *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. pp. 9621–9630 (2019)
31. Xiong, G., Wang, Y., Li, Z., Yang, W., Zhang, T., Zhou, X., Zhang, S., Yongdong, Z.: Aggregation and purification: Dual enhancement network for point cloud few-shot segmentation. In: *IJCAI* (2024)
32. Ye, X., Li, J., Huang, H., Du, L., Zhang, X.: 3d recurrent neural networks with context fusion for point cloud semantic segmentation. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 403–417 (2018)
33. Zhang, C., Wu, Z., Wu, X., Zhao, Z., Wang, S.: Few-shot 3d point cloud semantic segmentation via stratified class-specific attention based transformer network. *arXiv preprint arXiv:2303.15654* (2023)
34. Zhang, Q., Wang, T., Hao, F., Wu, F., Cheng, J.: Prototype expansion and feature calibration for few-shot point cloud semantic segmentation. *Neurocomputing* **558**, 126732 (2023)

35. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16259–16268 (2021)
36. Zhao, N., Chua, T.S., Lee, G.H.: Few-shot 3d point cloud semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8873–8882 (2021)