Aggregation and Purification: Dual Enhancement Network for Point Cloud Few-shot Segmentation

Guoxin Xiong[1] , **Yuan Wang**[1] , **Zhaoyang Li**[1] , **Wenfei Yang**[1] , **Tianzhu Zhang**[1,2,3*] , **Xu Zhou**[4] , **Shifeng Zhang**[4] and **Yongdong Zhang**[1,2]
[1]University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, China
[3]Deep Space Exploration Lab
[4]Sangfor Technologies Inc
{xgx, wy2016, lizhaoyang, yangwf}@mail.ustc.edu.cn, tzzhang@ustc.edu.cn, {zhouxu, zhangshifeng}@sangfor.com.cn, zhyd73@ustc.edu.cn

Abstract

Point cloud few-shot semantic segmentation (PC-FSS) aims to segment objects within query samples of new categories given only a handful of annotated support samples. Although PC-FSS demonstrates enhanced category generalization capabilities compared to the fully supervised paradigm, the prevalent significant scene discrepancies, which can be systematically summarized into intra-semantic diversity and semantic inconsistency, have posed substantial challenges to the area. In this work, we design a novel Dual Enhancement Network (DENet) to comprehensively tackle different kinds of scene discrepancies in a coherent and synergistic framework. The proposed DENet enjoys several merits. First, we design a mutual aggregation module to reconcile the intrinsic tension between the support prototypes and query point features, and the intra-semantic diversity is diminished in a bidirectional manner. Second, the consistent purification strategy is introduced to eliminate ambiguous prototypes, thereby reducing the mismatches brought by semantic inconsistency. Extensive experiments on S3DIS and ScanNet under different settings demonstrate that DENet significantly outperforms previous SOTAs.

1 Introduction

Point cloud semantic segmentation, a fundamental task in computer vision with wide applications in autonomous driving, robotics, etc, has achieved conspicuous advancements with the development of deep learning [Long *et al.*, 2015; Sun *et al.*, 2023c; Wang *et al.*, 2024; Sun *et al.*, 2023d; Luo *et al.*, 2023; Mai *et al.*, 2024b]. This progress can be attributed to well-established datasets and elaborately designed algorithms. However, the inherent category sensitivity of the fully supervised segmentation models restricts their capacity to predefined training categories. In pursuit of the human-like intelligence of recognizing new classes with only a glance, point cloud few-shot segmentation [Zhao *et al.*, 2021b] (PC-FSS) is proposed to extend the segmentation model to novel categories without extensive labeled data collection and time-consuming model retraining. Specifically, PC-FSS achieves the segmentation of query point clouds in new classes, which were previously unseen during the training process, using only a handful of annotated support samples.

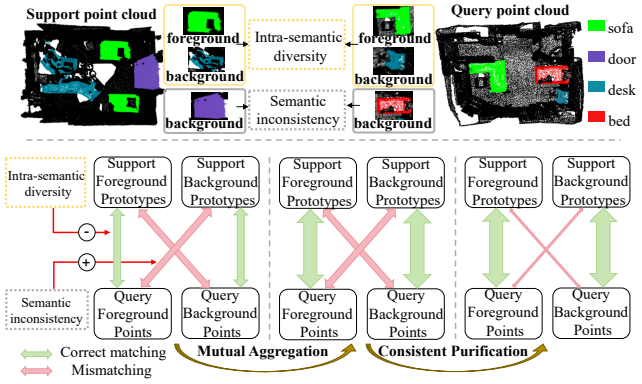

Figure 1: In PC-FSS, the presence of intra-class diversity and semantic inconsistency create bottlenecks in the area. We introduce the mutual aggregation module to alleviate intra-semantic diversity from both the support and query aspects, and the consistent purification strategy to mitigate the impact of semantic-inconsistency.

In previous literature, prototypical learning has emerged as the mainstream paradigm to tackle the PC-FSS. In specific, these methods [Ning *et al.*, 2023; Zhao *et al.*, 2021b] concentrate information of the support samples into a set of representative prototypes to guide the classification of query points. Despite yielding promising results, the frequently encountered significant scene discrepancies result in the prototypes derived solely from support samples struggling to furnish query points classification with essential cues.

As one of the principal challenges of PC-FSS, the discrepancies between the support and query samples can be systematically classified into the following two types: **intra-semantic diversity** and **semantic inconsistency**. The former

*Corresponding author

primarily originates from categories that concurrently appear in both support and query samples. (As the foreground category "sofa" and background category "desk" in Figure 1). Although the support prototypes containing co-occurrent information are capable of aiding in the segmentation of the query counterparts, they simultaneously risk introducing cognitive biases due to intra-class diversity such as variety of scale, pose, and so on. While the semantic inconsistency primarily exists in backgrounds including distinct semantics (As the "door" and "bed" in Figure 1). The resulting inconsistent support prototypes may be redundant or even disruptive to the classification of query points, and the query background points with the semantics that are absent in support samples may also be erroneously assigned to the foreground prototypes. To make matters worse, the negative impact of scene discrepancies is inevitably amplified by inbuilt low-data regimes of few-shot segmentation, leading to suboptimal results.

A series of works attempt to resolve the issue by calibrating the support prototypes with query context [Ning *et al.*, 2023; Zhu *et al.*, 2023; Zhang *et al.*, 2023]. While these methods can improve the query-awareness of prototypes in certain scenarios, the strategy of indiscriminately aggregating query information without accounting for different types of discrepancies tends to be highly sensitive to variations in scene composition. On the other hand, existing methodologies predominantly focus on unidirectional updating support prototypes, overlooking the crucial aspect that query features, as the primary entities of segmentation, also necessitate enhancement to be more task-aware.

After an in-depth analysis of the above issues, we design a novel Dual Enhancement Network (**DENet**), including a mutual enhancement aggregation and a consistent purification strategy, to coherently tackle different kinds of scene discrepancies by reconciling the intrinsic tension between the support prototypes and query point features in a bidirectional manner. **To deal with the intra-semantic diversity**, in the mutual aggregation module (MAM), we enable bidirectional feature communication between support prototypes and query features via a sequence of customized attention layers. Compared to the vanilla cross-attention adopted in previous works [Ning *et al.*, 2023], the feature interaction within MAM not only reduces intra-semantic diversity from both support and query aspects, but also endows the primary entities of segmentation, i.e., the query features with task-awareness and clean separation of background/foreground semantics. **To deal with the semantic inconsistency**, we elegantly design the consistent purification strategy (CPS) to explicitly eliminate ambiguous prototypes by comparing them with robust holistic features. Then the refined prediction from the filtered prototypes is adopted as the pseudo label to impose an additional regularization loss, which encourages the aggregation process of MAM to concentrate more on semantic consistent features. Moreover, the CPS can be employed iteratively, and its inherently parameter-free nature enables it to function as a plug-and-play module in the inference phase.

To recap, our primary contributions are concluded as follows: (1) We propose the Dual Enhancement Network (DENet) to tackle different types of scene discrepancies in a coherent and synergistic framework. (2) We introduce the mutual aggregation module (MAM) to alleviate intra-semantic diversity from both the support and query aspects, and the plug-and-play consistent purification strategy (CPS) to mitigate the impact of semantic-inconsistency. (3) Extensive experiments show that DENet significantly outperforms previous SOTAs. The systematic analysis of scene discrepancies in PC-FSS also sheds light on future research from a more comprehensive perspective.

2 Related Work

3D Point Cloud Semantic Segmentation. With the recent advances in deep neural networks [Sun *et al.*, 2021; Landrieu and Simonovsky, 2018; Sun *et al.*, 2023a; Mai *et al.*, 2023; Sun *et al.*, 2023b; Wang *et al.*, 2023a; Luo *et al.*, 2024; Pan *et al.*, 2023; Mai *et al.*, 2024a; Xiong *et al.*, 2024], point cloud semantic segmentation has gained significant attention as a solution to address the challenges posed by dense pixel-level annotations. Recently, many deep learning based approaches [Hu *et al.*, 2020; Huang *et al.*, 2018; Lai *et al.*, 2022; Wu *et al.*, 2019; Li *et al.*, 2018; Qi *et al.*, 2017b; Wang *et al.*, 2019; Ye *et al.*, 2018; Zhao *et al.*, 2021a] are proposed to address 3D point cloud semantic segmentation under fully supervised conditions. PointNet [Qi *et al.*, 2017a] introduces a pioneering point-based method to segment raw point clouds directly and extract point features with MLPs. Despite its simplicity and efficiency, PointNet falls short in capturing essential local information from neighboring points. DGCNN [Wang *et al.*, 2019] introduces EdgeConv to address this issue, capturing local point cloud structures through a dynamic k-NN graph for long-distance point intersection analysis. In this work, DGCNN serves as the backbone of our feature extractor to extract local structural characteristics and semantic features. Building upon this, we introduce a novel approach aimed at addressing the challenges of 3D point cloud semantic segmentation in a few-shot learning context.

Point Cloud Few-shot Semantic Segmentation. Few-shot 3D Point Cloud Semantic Segmentation transitions traditional 3D point cloud semantic segmentation into a few-shot context [Wang *et al.*, 2022; Wang *et al.*, 2023b], equipping the model with the capability to segment novel classes using minimal support data. AttMPTI [Zhao *et al.*, 2021b] innovatively applied an attention-aware multi-prototype transductive inference approach to few-shot 3D point cloud semantic segmentation, establishing a foundational method in this domain for classifying new categories using a limited number of annotated examples. However, AttMPTI did not fully address the gap issue between support and query. Subsequent works [Ning *et al.*, 2023; Zhu *et al.*, 2023; Zhang *et al.*, 2023] have utilized information from the query to enhance the adaptability of support prototypes to the context of the query. Building upon these prior efforts, our method takes a step further by addressing the discrepancies between support and query from two perspectives: intra-semantic diversity and semantic inconsistency.

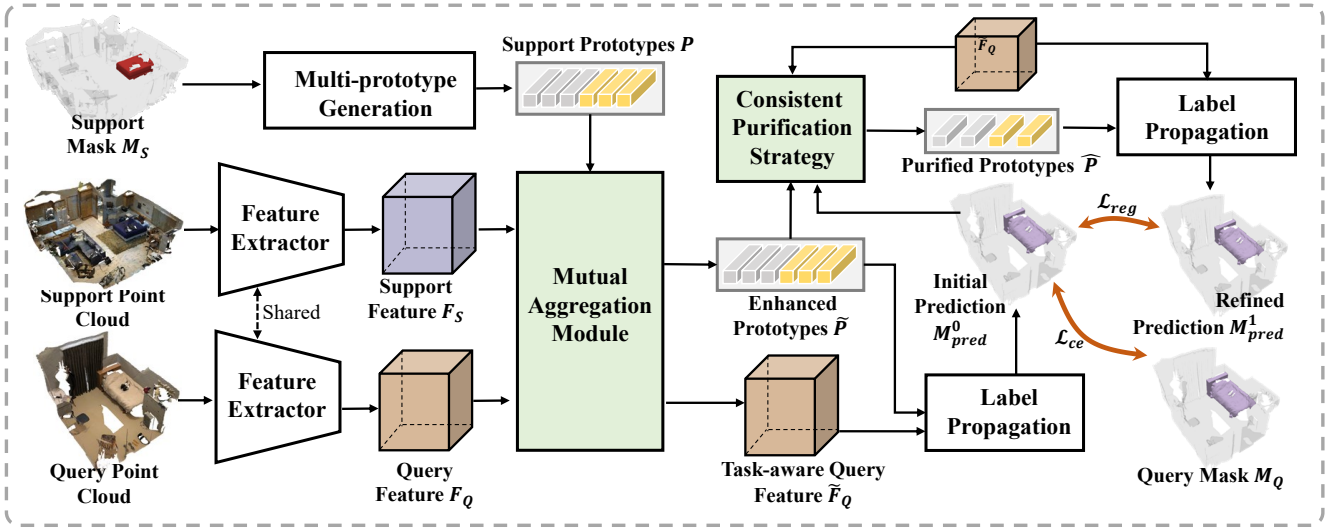

Figure 2: The architecture of the proposed Dual Enhancement Network (DENet). We introduce the mutual aggregation module (MAM) and the consistent purification strategy (CPS) to tackle different types of scene discrepancies in a coherent and synergistic framework.

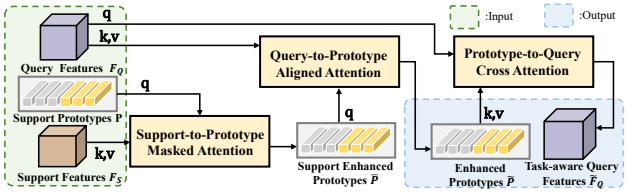

Figure 3: Illustration of the mutual aggregation module.

3 Method

3.1 Problem Definition

Point cloud few-shot semantic segmentation aims to endow the segmentation model with the ability to rapidly generalize to novel categories with only a few set of labeled samples. Following previous works [Zhao *et al.*, 2021b; Ning *et al.*, 2023], we adopt the widely used episodic meta-training paradigm [Vinyals *et al.*, 2016]. Specifically, considering two non-intersection sets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, which possess non-overlapping target categories ($\mathcal{C}_{\text{train}} \cap \mathcal{C}_{\text{test}} = \emptyset$). A series of episodes are sampled from $\mathcal{D}_{\text{train}}$ to train our model, and then we test the trained model on the test set $\mathcal{D}_{\text{test}}$. Each episode, manifests as an N-way K-shot segmentation task within point cloud data. And each episode is composed of a support set denoted as $\mathcal{S} = \{(I_S^{n,k}, M_S^{n,k})\}$, where I is point cloud, M is mask, $k \in \{1, \cdots, K\}$, $n \in \{1, \cdots, N\}$, and a query set $Q = \{(I_Q, M_Q)\}$. During each episode in N-way K-shot scenario, the model is provided with input data comprising $\{I_S^{n,k}, M_S^{n,k}, I_Q\}$, where $I_S^{n,k}$ and $M_S^{n,k}$ represent the point cloud and mask of the support set, respectively, and I_Q represents the query point cloud. The model predict the segmentation result for I_Q, which is compared to M_Q as the ground truth during validation. During the training phase, the model is trained to infer the segmentation of the query set I_Q by leveraging the information provided in the support set S, guided by the ground truth mask M_Q. In the testing phase, the generalization capabilities of the model are assessed using tasks derived from D_{test}. To simplify the description, we describe our approach in a 1-way 1-shot setting.

3.2 Overview of Dual Enhancement Network

As illustrated in Figure 2, the proposed DENet follows the prototypical learning paradigm and mainly focuses on dealing with the misalignment between the support prototypes $\mathrm{P} = \{(\mathrm{P_S}^{bg}, \mathrm{P_S}^{fg})\}$ originate from support point features $\mathrm{F_S} \in \mathbb{R}^{N \times C}$ and query point features $\mathrm{F_Q} \in \mathbb{R}^{N \times C}$. Note that we adopt the same prototype generation and transductive inference process as done in [Zhao *et al.*, 2021b] for a fair comparison without claiming any contribution. DENet synergistically integrates two components, *i.e.*, Mutual Aggregation Module (MAM) and Consistent Purification Strategy (CPS), to handle different types of discrepancies. MAM enables the bidirectional feature aggregation to bridge the intra-semantic gap while preserving the foreground-background discriminability (Section 3.3). CPS mitigates the negative impact of semantic inconsistency by eliminating ambiguous prototypes. (Section 3.4). The details are as follows.

3.3 The Mutual Aggregation Module

Due to the significant intra-class diversity between query and support samples, the local prototypes derived solely from support samples fail to comprehensively capture class information, thereby proving inefficient as directors for the classification of query points. We design the MAM to promote a more holistic object pattern understanding of both the support prototypes and query features, thus reducing the impact of the intra-semantic gap. MAM consists of three different attention modules, namely the support-to-prototype masked attention module, the query-to-prototype alignment attention module, and the prototype-to-query attention module.

Support-to-Prototype Masked Attention. Since prototypes are obtained by averaging small portions of point features [Zhao *et al.*, 2021b], only local and limited information is contained. we adopt masked attention layers to aggregate support features to the corresponding foreground or background prototypes to endow them with a holistic perception of the category, such as geometric structures. The standard cross-attention process is formulated as

$$\overline{\mathrm{P}} = \mathrm{Softmax}\left(\frac{\mathrm{W_q(P)W_k(F_S)}}{\sqrt{\mathrm{C}}}\right)\mathrm{W_v(F_S)} + \mathrm{P}, \quad (1)$$

where $\sqrt{C}$ is a scaling factor, $\mathrm{W_q}, \mathrm{W_k}, \mathrm{W_v}$ are linear projections of query, key, and value in the same dimension C. However, the soft nature of attention makes the process noisy and less dependable - foreground prototypes might have small weights distributed in the background features and vice versa. Inspired by [Cheng *et al.*, 2022], we deploy masked attention to aid the clean separation of semantics between foreground and background. In practice, we force foreground prototypes to focus on the foreground features of the support sample while directing background prototypes to interact with the background support features, formally:

$$\overline{\mathrm{P}} = \mathrm{Softmax}\left(\frac{\mathrm{W_q(P)W_k(F_S)}}{\sqrt{\mathrm{C}}} + \mathcal{M}_{\mathrm{sup}}\right)\mathrm{W_v(F_S)} + \mathrm{P}, \quad (2)$$

where $\mathcal{M}_S \in \{0, -\infty\}^{np \times N}$ determines whether the query p is allowed($= 0$) or not allowed ($= -\infty$) to attend to the i-th feature, defined as:

$$\mathcal{M}_{sup}(p,i) = \begin{cases} 0 & \text{if } p \in \mathrm{P_S^{fg}} \text{ and } M_S(i) = 1 \\ 0 & \text{if } p \in \mathrm{P_S^{bg}} \text{ and } M_S(i) = 0 \\ -\infty & \text{otherwise} \end{cases}. \quad (3)$$

In this way, $\mathrm{P_S^{fg}}$ and $\mathrm{P_S^{bg}}$ are restricted to absorbing contextual information from the foreground and background respectively. The resulting prototypes $\overline{\mathrm{P}}$, equipped with task awareness, also effectively avoid cross-category interference.

Query-to-Prototype Alignment Attention. To enhance the semantics perception of prototypes from the query perspective, thereby reducing the gap between them, we aggregate query category knowledge into $\overline{P}$ via cross-attention layers. Considering that the query mask is not accessible, we introduce an alignment matrix $\mathcal{A}$ to filter out implausible matching. Formally:

$$\tilde{\mathrm{P}} = \mathrm{Softmax}\left(\frac{\mathrm{W_q}(\overline{\mathrm{P}})\mathrm{W_k(F_Q)}}{\sqrt{\mathrm{C}}} + \mathcal{A}\right)\mathrm{W_v(F_Q)} + \overline{\mathrm{P}}, \quad (4)$$

where $\sqrt{C}$ is a scaling factor, $\mathrm{W_q}, \mathrm{W_k}, \mathrm{W_v}$ are linear projections of query, key, and value in the same dimension C, same as before but with different weights. The alignment matrix $\mathcal{A}$ is obtained by

$$\mathcal{A}(i,j) = \begin{cases} 0 & \text{if } i \in \mathrm{P_S^{fg}} \text{ and } argmax\langle j, \overline{\mathrm{P}}\rangle \in \mathrm{P_S^{fg}} \\ 0 & \text{if } i \in \mathrm{P_S^{bg}} \text{ and } argmax\langle j, \overline{\mathrm{P}}\rangle \in \mathrm{P_S^{bg}} \\ -\infty & \text{otherwise} \end{cases}. \quad (5)$$

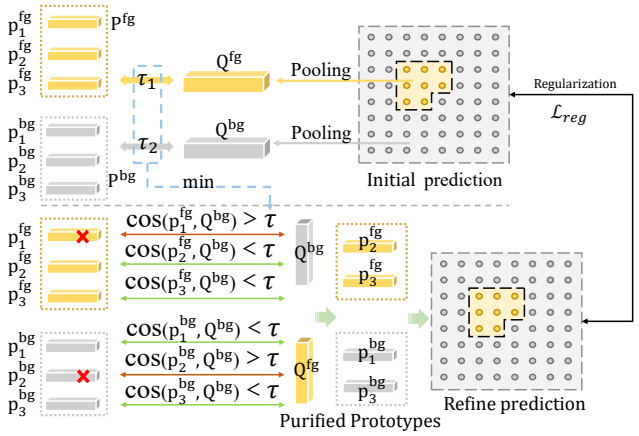

Figure 4: Illustration of the Consistent Purification Strategy(CPS), which mitigates the semantic inconsistency by evaluating and filtering the support prototypes to increase the reliability and accuracy of the query point classification.

The alignment matrix ensures the correspondence in feature aggregation, and the query-to-prototype alignment attention significantly reduces the intra-semantic diversity between the prototype and query features.

Prototype-to-Query Attention. In addition to modulating the support prototypes to approach query samples, we deem that the query features, as the main entity for the segmentation task, can also be calibrated toward the support prototypes, forming a bidirectional alignment. To achieve that, prototype-to-query attention is included to aggregate discriminative prototype features. Specifically, we employ cross-attention with residual connections to update query features with prototypes:

$$\tilde{\mathrm{F}}_{\mathrm{Q}} = \mathrm{Softmax}\left(\frac{\mathrm{W_q(F_Q)W_k}(\tilde{\mathrm{P}})}{\sqrt{\mathrm{C}}}\right)\mathrm{W_v}(\tilde{\mathrm{P}}) + \mathrm{F_Q}. \quad (6)$$

The reversed cross-attention layer injects task-related information into query features, enabling a more consistent matching process for label prediction.

3.4 The Consistent Purification Strategy

Similar to AttMPTI [Zhao *et al.*, 2021b], we use the graph relationship between bi-directional enhanced support prototypes $\tilde{\mathrm{P}}$ and query features $\tilde{\mathrm{F}}_{\mathrm{Q}}$ to derive the initial prediction results for each point in the query sample using a label propagation algorithm, denoted as $M^0_{pred} \in \{0,1\}^{1\times N}$.

However, due to the presence of semantic inconsistency, some of the support prototypes may be redundant or even disruptive to the label propagation. That is, some background points of the query may inappropriately correspond to the support foreground prototypes, and vice versa. Therefore, we introduce a consistent purification strategy, which evaluates and filters the support prototypes to reduce the negative impact of semantic inconsistency.

To delve into the correlation between support prototypes and query samples, we first apply mask average pooling (MAP) to the initial prediction of query features M^0_{pred} to

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	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	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	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	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	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	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	**75.99**	**78.24**	**77.11**	**82.57**	**84.11**	**83.34**	**59.92**	**61.88**	**60.90**	**67.34**	74.23	**70.78**

Table 1: 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	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	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	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	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	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	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	**62.99**	**64.92**	**63.95**	**68.35**	**73.92**	**71.13**	**44.25**	**48.57**	**46.41**	**53.24**	**58.86**	**56.05**

Table 2: 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**

obtain the holistic foreground and background information of the query, denoted as Q^{fg} and Q^{bg}, respectively.

$$\mathbf{Q^{fg}} = \frac{\sum_{i=1}^{N} M^0_{pred}(i) \cdot \tilde{\mathrm{F}_{\mathrm{Q}}}(i)}{\sum_{i=1}^{N} M^0_{pred}(i)}, \tag{7}$$

$$\mathbf{Q^{bg}} = \frac{\sum_{i=1}^{N} (1 - M^0_{pred}(i)) \cdot \tilde{\mathrm{F}_{\mathrm{Q}}}(i)}{\sum_{i=1}^{N} (1 - M^0_{pred}(i))}. \tag{8}$$

Then, by measuring the average correlation between the support foreground prototypes $\tilde{\mathrm{P}}^{\mathrm{fg}}$ and the holistic query foreground feature Q^{fg}, denoted as τ_1, as well as the average correlation between the support background prototypes $\tilde{\mathrm{P}}^{\mathrm{bg}}$ and the holistic query background feature Q^{bg}, denoted as τ_2. We take the minimum of τ_1 and τ_2 as the holistic relevance of the support prototypes to the query samples, which serves as the prototype purification threshold τ, *i.e.*$\tau = \min(\tau_1, \tau_2)$.

$$\begin{aligned} \mathbf{R^{sb-qf}} &= \cos(\tilde{\mathrm{P}}^{\mathrm{bg}}, \mathrm{Q}^{\mathrm{fg}}) \quad \mathbf{R^{sb-qb}} = \cos(\tilde{\mathrm{P}}^{\mathrm{bg}}, \mathrm{Q}^{\mathrm{bg}}), \\ \mathbf{R^{sf-qb}} &= \cos(\tilde{\mathrm{P}}^{\mathrm{fg}}, \mathrm{Q}^{\mathrm{bg}}) \quad \mathbf{R^{sf-qf}} = \cos(\tilde{\mathrm{P}}^{\mathrm{fg}}, \mathrm{Q}^{\mathrm{fg}}), \end{aligned} \tag{9}$$

$$\tau_1 = \mathbf{AVG}(\mathbf{R^{sf-qf}}) \qquad \tau_2 = \mathbf{AVG}(\mathbf{R^{sb-qb}}), \tag{10}$$

where the "AVG" denotes the average operation across all prototypes. We base this on the simple but straightforward assumption that the foreground-to-foreground similarity and background-to-background similarity should be higher than the background-to-foreground or foreground-to-background similarity.

$$\mathbf{C}(i) = \begin{cases} 1 & \text{if } i \in \tilde{\mathrm{P}}^{\mathrm{bg}} \text{ and } \cos(i, \mathrm{Q}^{\mathrm{fg}}) < \tau \\ 1 & \text{if } i \in \tilde{\mathrm{P}}^{\mathrm{fg}} \text{ and } \cos(i, \mathrm{Q}^{\mathrm{bg}}) < \tau, \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where $\mathbf{C}(i)$ is a judgment function on the i^{th} prototype that decides whether to keep P^i ($\mathbf{C}(i)$=1) or to filter out the inconsistent P^i ($\mathbf{C}(i)$=0). The purified prototypes $\hat{\mathrm{P}}$ are subsequently used in label propagation along with the query features $\tilde{\mathrm{F}}_{\mathrm{Q}}$ to get the refined prediction M^1_{pred}. In addition to the baseline design cross-entropy loss $\mathcal{L}_{ce}$ for initial query prediction M^0_{pred}, the CPS introduces an additional regularization loss to encourage the aggregation process of MAM to focus more on semantically consistent features,

$$\mathcal{L}_{reg} = \mathit{CrossEntropy}(M^0_{pred}, M^1_{pred}), \tag{12}$$

and the overall training objective $\mathcal{L}$ is defined as:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{reg}. \tag{13}$$

It should be noted that the CPS not only introduces supplementary regularization loss during the training phase, but also acts as a plug-and-play module to refine the prediction.

4 Experiments

4.1 Experiment Setup

Datasets. In accordance with previous methods, we evaluate the proposed DANet on two commonly 3D few-shot

MEM	CPS	mIoU
		66.27
✓		72.17
	✓	71.33
✓	✓	**75.99**

Table 3: Ablation study on different components.

S2P	Q2P	P2Q	mIoU
			71.33
✓			72.65
	✓		72.78
✓	✓		74.51
✓	✓	✓	**75.99**

Table 4: Ablation study on mutual aggregation module components.

Train	Infer.	mIoU
		72.17
✓		74.11
	✓	73.84
✓	✓	**75.99**

Table 5: Deployment of consistent purification strategy.

Iter.	mIoU
1	73.84
2	74.15
3	**74.38**
4	74.38
5	74.38

Table 6: The results of iterative inference.

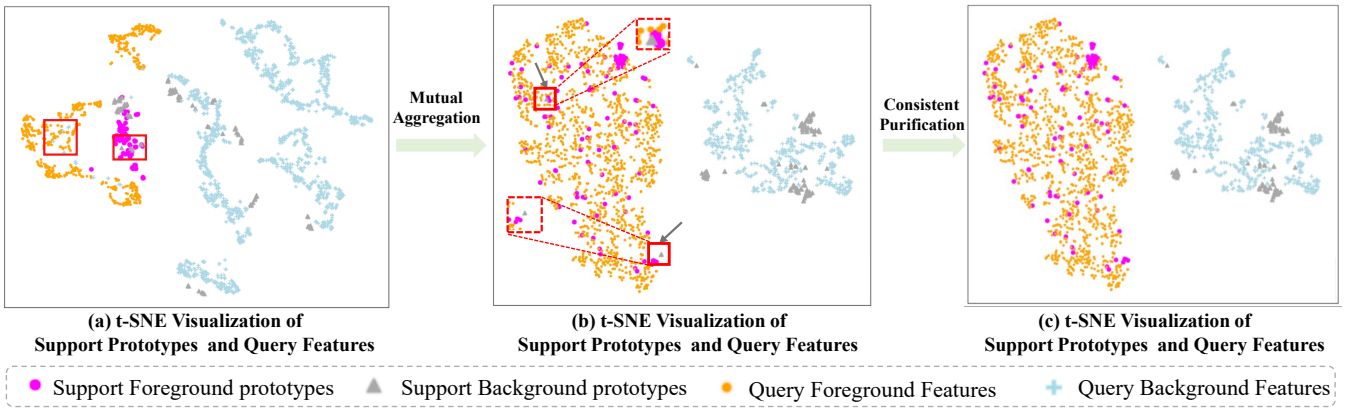

Figure 5: t-SNE visualization of two aspects of scene discrepancies, *i.e.*, intra-semantic diversity and semantic inconsistency, and the process by which the DENet mitigates the discrepancies. More details refer to "**Investigation of Dual Enhancement Network**".

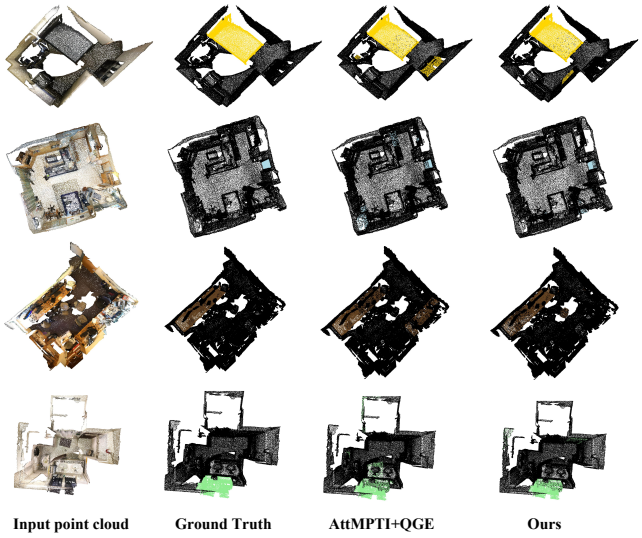

Figure 6: Qualitative results of our method in comparison to the ground truth and AttMPTI+QGE. The target classes from top to bottom are "*bed*" (first row, yellow), "*door*" (second row, blue), "*desk*" (third row, brown), and "*window*" (last row, green).

segmentation benchmarks, **S3DIS** [Armeni *et al.*, 2016] and **ScanNet** [Dai *et al.*, 2017]. S3DIS contains 272 point cloud datasets covering indoor areas from three different buildings. It includes a wide range of indoor environments such as lobbies, corridors, offices, and pantries. Each point cloud dataset is annotated with 12 different semantic categories, along with an additional background category to represent various surrounding clutter. ScanNet consists of 1,513 point clouds extracted from 707 different indoor scenes, and provides a rich annotation framework with 20 semantic categories for segmentation and an additional class for unannotated regions. Compared to S3DIS, ScanNet presents a wider variety of room types, including bathroom and living room, with more irregular point cloud scenes. In each dataset, the classes are systematically partitioned into two mutually exclusive subsets, denoted as $\mathcal{S}_0$ and $\mathcal{S}_1$, ensuring no overlap. These subsets are alternately utilized for cross-validation purposes, with one serving as the test classes $\mathcal{C}_{\text{test}}$ and the other as the training classes $\mathcal{C}_{\text{train}}$. Consistent with the preprocessing strategy of previous work, we divide rooms into 1m x 1m blocks, with 2048 points randomly sampled for each block.

Evaluation Metric. Following previous works [Zhao *et al.*, 2021b; Ning *et al.*, 2023], we adopt mean-IoU as our evaluation metrics. In our few-shot setting, the mean-IoU is obtained by averaging the IoU over all testing classes $\mathcal{C}_{\text{test}}$.

Implementation Details. All models are implemented using the PyTorch framework and trained 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. We use the Adam optimizer with a learning rate of 0.001 to optimize the pre-trained model. During episodic training, we initialize the feature extraction module by loading the pre-trained weights. The initial learning rate is set to 0.0001 for the feature extraction module and 0.001 for

the other modules, and the Adam optimizer is used to update all parameters. The learning rates are halved every 5000 iterations and the batch size is set to 1. Following [Zhao *et al.*, 2021b; Ning *et al.*, 2023], models undergo 40,000 iterations. During each iteration, we randomly sample an episode, utilizing Gaussian noise and arbitrary z-axis rotations to augment the point clouds in both the support and query sets.

4.2 Comparison with State-of-the-Art Methods

We experimented on two popular benchmarks, S3DIS and ScanNet, comparing our method with previous PC-FSS methods in Table 1 and Table 2, respectively. Across various evaluation scenarios on both benchmarks, the proposed DENet consistently outperforms prior advanced approaches, strongly proving the effectiveness of our method. Specifically, on ScanNet, our DENet achieves 63.95% and 71.13% mIoU for 1-way 1-shot and 1-way 5-shot settings. This performance represents an improvement of 3.59% and 1.6% over the most competitive approach, QGE [Ning *et al.*, 2023]. We additionally observed that the enhancements offered by our proposed DENet are more pronounced on the more complex ScanNet dataset as compared to the S3DIS dataset. This can be attributed to the greater complexity of scenes and categories present in the ScanNet dataset, leading to more conspicuous discrepancies between support and query. Consequently, a series of results demonstrate that our DENet, developed through a systematic analysis of scene discrepancies in the PC-FSS task, exhibits commendable performance and robustness in handling complex scene scenarios. Figure 6 displays prediction visualizations on ScanNet dataset, illustrating that our proposed method yields more visually appealing results, intuitively demonstrating the superiority of our approach.

4.3 Ablation Study

In this section, we use AttMPTI as our baseline model and conduct extensive ablation studies on S3DIS S^0 split under the 1-way 1-shot setting to verify the effectiveness of each component of our DENet, including the mutual enhancement module and the consistent purification strategy.

Effectiveness of Components. As shown in Table 3, both mutual aggregation module (MAM) and consistent purification strategy (CPS) consistency bring a certain performance lift compared with the baseline. (1) With the utilization of MAM, a 5.9% improvement of mIoU can be observed, indicating that reducing intra-semantic diversity between support and query samples can benefit the point cloud few segmentation task. (2) The introduction of CPS achieves further accuracy gains (3.8% in mIoU), primarily attributed to the exclusion of prototypes with ambiguous semantics, which helps to suppress the negative impact of scene discrepancies.

Effectiveness of Mutual Aggregation Module. To perform ablation studies on the components of our Mutual Aggregation module, we used the AttMPTI model as our baseline, while all experiments used the CPS. As shown in Table 4, both the Prototype-Support Masked Attention module and the Prototype-Query Alignment Attention module can improve the accuracy, indicating that optimizing support prototypes can better align them with the query samples. This is consistent with previous methods of updating support prototypes to improve segmentation performance. By further integrating the Query-Prototype Attention module, we observe a performance advance of 1.48% mIoU, attributed to the module's ability to extract consistent information from the query perspective and mitigate the negative effects of the semantic gaps between the query and support point clouds.

Effectiveness of Consistent Purification Strategy. To further explore the applicability of CPS, as described in Table 5, we examine its effectiveness in both the training and testing phases. The introduce of CPS consistently leads to significant performance improvements in both training and testing scenarios. Furthermore, the use of CPS during both training and testing further elevates performance. This is because CPS effectively eliminates ambiguous prototypes and reducing the impact of category inconsistency on model performance.

Hyperparameter Evaluations. Quantitative experiments are conducted to explore the appropriate number of iterations. In Table 6, we compare the performance of different iterations of the CPS. It can be observed that as the number of purification iterations increases, the performance gradually increases and then converges. To achieve a trade-off between accuracy and speed, we finally choose a number of iterations of 3.

Investigation of Dual Enhancement Network. In Figure 5, we employed t-SNE visualizations to concretely illustrate two types of scene discrepancies, as well as the process by which the DENet mitigates them. The intra-semantic diversity can be observed from the orange dots and magenta dots in Figure 5 (a). They represent the foreground features of support and query points belonging to the same category but are not well aligned. The semantic inconsistency is primarily reflected in the intertwining of background interference features with foreground features as shown by the darkgray triangles and lightblue crosses within the red boxes in Figure 5 (a). In the presence of scene discrepancies, support prototypes struggle to effectively guide the classification of query features. The integration of the mutual aggregation module (MAM) markedly enhances the intra-class alignment, we deem this is reasonable as the adaptive feature communication within the customized markedly alleviates intra-semantic diversity in a bidirectional manner while preserving the foreground-background discriminability of corresponding features. The consistent purification strategy (CPS) further eliminates ambiguous prototypes and thus effectively suppresses erroneous label propagation. The MAM and CPS synergistically reinforce each other, thereby constituting the coherent PC-FSS framework DENet, which is capable of addressing different types of scene discrepancies.

5 Conclusion

In this paper, we propose a novel Dual Enhancement Network (DENet) for point cloud few-shot semantic segmentation. By systematically analyze the scene discrepancies in the PC-FSS task, we tailored the MAM module and the CPS module to respectively address intra-semantic discrepancies and semantic inconsistencies issues. Extensive experiments demonstrate the effectiveness of our method.

Contribution Statement

Guoxin Xiong and Yuan Wang contributed equally to this paper.

Acknowledgments

This work was partially supported by the National Defense Basic Scientific Research Program (Grant JCKY20211130B016).

References

[Armeni *et al.*, 2016] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016.

[Cheng *et al.*, 2022] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

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

[Hu *et al.*, 2020] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11108–11117, 2020.

[Huang *et al.*, 2018] Qiangui Huang, Weiyue Wang, and Ulrich Neumann. Recurrent slice networks for 3d segmentation of point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2626–2635, 2018.

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

[Landrieu and Simonovsky, 2018] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4558–4567, 2018.

[Li *et al.*, 2018] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[Luo *et al.*, 2023] Naisong Luo, Yuwen Pan, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Camouflaged instance segmentation via explicit de-camouflaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17927, 2023.

[Luo *et al.*, 2024] Naisong Luo, Rui Sun, Yuwen Pan, Tianzhu Zhang, and Feng Wu. Electron microscopy images as set of fragments for mitochondrial segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[Mai *et al.*, 2023] Huayu Mai, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Dualrel: Semi-supervised mitochondria segmentation from a prototype perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19617–19626, 2023.

[Mai *et al.*, 2024a] Huayu Mai, Rui Sun, Yuan Wang, Tianzhu Zhang, and Feng Wu. Pay attention to target: Relation-aware temporal consistency for domain adaptive video semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[Mai *et al.*, 2024b] Huayu Mai, Rui Sun, Tianzhu Zhang, and Feng Wu. Rankmatch: Exploring the better consistency regularization for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[Ning *et al.*, 2023] Zhenhua Ning, Zhuotao Tian, Guangming Lu, and Wenjie Pei. Boosting few-shot 3d point cloud segmentation via query-guided enhancement. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1895–1904, 2023.

[Pan *et al.*, 2023] Yuwen Pan, Naisong Luo, Rui Sun, Meng Meng, Tianzhu Zhang, Zhiwei Xiong, and Yongdong Zhang. Adaptive template transformer for mitochondria segmentation in electron microscopy images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21474–21484, 2023.

[Qi *et al.*, 2017a] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[Qi *et al.*, 2017b] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[Sun *et al.*, 2021] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware

transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10938–10947, 2021.

[Sun *et al.*, 2023a] Rui Sun, Naisong Luo, Yuwen Pan, Huayu Mai, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Appearance prompt vision transformer for connectome reconstruction. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 1423–1431. International Joint Conferences on Artificial Intelligence Organization, 8 2023. Main Track.

[Sun *et al.*, 2023b] Rui Sun, Huayu Mai, Naisong Luo, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Structure-decoupled adaptive part alignment network for domain adaptive mitochondria segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 523–533. Springer, 2023.

[Sun *et al.*, 2023c] Rui Sun, Huayu Mai, Tianzhu Zhang, and Feng Wu. Daw: Exploring the better weighting function for semi-supervised semantic segmentation. In *Advances in Neural Information Processing Systems*, 2023.

[Sun *et al.*, 2023d] Rui Sun, Yuan Wang, Huayu Mai, Tianzhu Zhang, and Feng Wu. Alignment before aggregation: trajectory memory retrieval network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1218–1228, 2023.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[Wang *et al.*, 2019] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019.

[Wang *et al.*, 2022] Yuan Wang, Rui Sun, Zhe Zhang, and Tianzhu Zhang. Adaptive agent transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 36–52. Springer, 2022.

[Wang *et al.*, 2023a] Yuan Wang, Naisong Luo, and Tianzhu Zhang. Focus on query: Adversarial mining transformer for few-shot segmentation. In *Advances in Neural Information Processing Systems*, 2023.

[Wang *et al.*, 2023b] Yuan Wang, Rui Sun, and Tianzhu Zhang. Rethinking the correlation in few-shot segmentation: A buoys view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2023.

[Wang *et al.*, 2024] Yuan Wang, Rui Sun, Naisong Luo, Yuwen Pan, and Tianzhu Zhang. Image-to-image matching via foundation models: A new perspective for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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

[Xiong *et al.*, 2024] Guoxin Xiong, Meng Meng, Tianzhu Zhang, Dongming Zhang, and Yongdong Zhang. Reference-aware adaptive network for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2024.

[Ye *et al.*, 2018] Xiaoqing Ye, Jiamao Li, Hexiao Huang, Liang Du, and Xiaolin Zhang. 3d recurrent neural networks with context fusion for point cloud semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 403–417, 2018.

[Zhang *et al.*, 2023] Qieshi Zhang, Tichao Wang, Fusheng Hao, Fuxiang Wu, and Jun Cheng. Prototype expansion and feature calibration for few-shot point cloud semantic segmentation. *Neurocomputing*, 558:126732, 2023.

[Zhao *et al.*, 2021a] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.

[Zhao *et al.*, 2021b] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Few-shot 3d point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8873–8882, 2021.

[Zhu *et al.*, 2023] Guanyu Zhu, Yong Zhou, Rui Yao, and Hancheng Zhu. Cross-class bias rectification for point cloud few-shot segmentation. *IEEE Transactions on Multimedia*, 2023.