

Generalized Few-Shot Point Cloud Segmentation via LLM-Assisted Hyper-Relation Matching

Zhaoyang Li^{1*} Yuan Wang^{1*} Guoxin Xiong^{1*} Wangkai Li¹ Yuwen Pan¹ Tianzhu Zhang^{1,2†}

¹University of Science and Technology of China

²National Key Laboratory of Deep Space Exploration, Deep Space Exploration Laboratory

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

Abstract

Generalized few-shot point cloud segmentation (GFS-3DSeg) aims to segment objects of both base and novel classes using abundant base class samples and limited novel class samples. Existing GFS-3DSeg methods encounter bottlenecks due to the scarcity of novel class data and inter-class confusion. In this paper, we propose the LLM-Assisted Hyper-Relation Matching (LARM) framework, which leverages the wealth of prior knowledge in Large Language Models (LLM) to enrich novel category prototypes and introduces a hyper-relation matching strategy to mitigate false matches between point features and category prototypes caused by inter-class confusion. The proposed LARM enjoys several merits. First, the vast knowledge embedded in LLM can be an effective complement to vanilla category prototypes, enabling them to exhibit greater robustness. Second, the hyper-relation matching strategy harnesses the structural information implicit in the inter-class relationships, making it more robust than individual feature comparisons. Extensive experiments on two benchmarks demonstrate that LARM outperforms previous state-of-the-art methods by large margins.

1. Introduction

Point cloud semantic segmentation is a fundamental computer vision task for scene understanding in autonomous driving, robotics, and virtual reality. Despite conspicuous achievements in fully supervised learning methods [9, 11, 13, 25, 26, 35, 45], they are limited to predefined training categories due to reliance on extensive annotated data. To address this issue, few-shot 3D point cloud semantic segmentation [46] (FS-3DSeg) has recently attracted increasing interest, aiming to derive segmentation models capable of generalizing to novel classes without laborious data

*Equal contribution

†Corresponding author

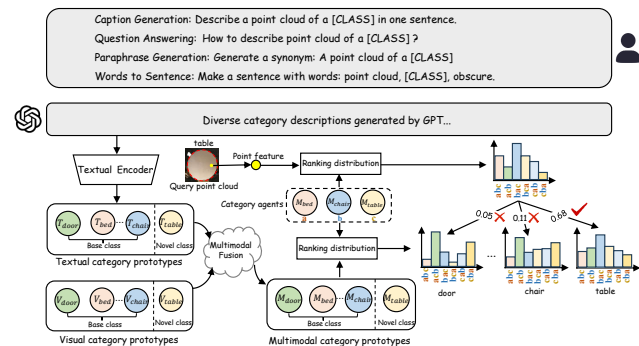


Figure 1. Illustration of our motivation. We leverage a wealth of knowledge of LLMs to compensate vanilla category representation and design a hyper-relation matching strategy for more robust point classification.

collection and model retraining. The top-performing FS-3DSeg methods [1, 10, 17, 19, 21, 34, 37, 42] adopt meta-training to capture transferable category correspondence from a few labeled support point clouds to novel category query point clouds. Though achieving remarkable progress, these FS-3DSeg methods are restricted to segmenting solely novel classes while ignoring base classes.

To address the above challenges, the generalized few-shot point cloud semantic segmentation [38] (GFS-3DSeg) is proposed to recognize both base and novel classes within a unified model during testing. Existing GFS-3DSeg methods [31, 38] typically adopt prototype learning paradigm [29], where base-class prototypes are first learned on abundant base data, and novel-class prototypes are subsequently generated from sparse support samples. These class prototypes are then matched with query point features to segment base and novel classes during inference.

Despite yielding promising results, these methods often encounter bottlenecks due to the scarcity and limited diversity of novel class data. On the one hand, novel category prototypes derived from sparse support samples often lack representativeness and cannot be comprehensive class descriptors [34], resulting in poor performance on novel classes [5]. On the other hand, the severe imbalance be-

tween the abundant base class data and the limited novel class samples can bias the learned feature representations towards the base classes. This can further exacerbate the intrinsic geometric ambiguity inherent in 3D point clouds, intensifying the inter-class confusion and making it even more challenging to distinguish between similar classes during the matching process, especially when novel categories (e.g., “table”) share similar shapes or contextual patterns with base classes (e.g., “chair”). Here, a critical question arises: ❶ *How can we improve the quality of novel class prototypes to enhance their representational and discriminative capabilities?*

Recent advancements in large language models (LLMs) demonstrate their potential to enhance performance for various downstream tasks [6, 8, 12, 27, 47–50]. Given that LLMs are pre-trained on large-scale corpora, they encompass a wealth of knowledge about the object’s semantics, structures, and contextual information. Inspired by this insight, we propose to harness LLMs (e.g., GPT-3 [3]) to empower GFS-3Dseg. In this way, the vast prior knowledge from LLMs can effectively compensate for the limitations of visual prototypes, particularly for novel classes. The resulting multimodal prototypes, fusing textual and visual information, exhibit enhanced robustness. Nonetheless, it is still non-trivial to precisely match these multimodal prototypes to their corresponding query point cloud features, as the inter-class confusion stemming from similar categories can potentially result in mismatches. Existing methods [31, 38] process each category independently during the matching phase, neglecting informative inter-class relations that implicitly capture structured contextual information for disambiguation. Here another question arises: ❷ *how to harness the structure information modeled in the inter-class relationship to facilitate more accurate matching?*

To address this question, we propose a novel hyper-relation similarity measurement that models the relations between a given point and a subset of semantically relevant category prototypes, termed as category agents. The key idea is to treat the category agent ranking as a probabilistic event rather than a deterministic permutation. Specifically, for a point p , the scores assigned to different categories can be interpreted as probabilities. These scores determine the ranking permutation, which reflects the relevance of each category to the point’s feature. An agent-ranking probability distribution can be constructed by associating the probability with every rank permutation for both the point feature and all category agents. As illustrated in Figure 1, we reformulate the similarity measurement from individual point-prototype comparisons to a more comprehensive agent-ranking hyper-relation similarity measurement. This approach effectively captures the structural information modeled in the inter-class relationships, enabling more accurate matching.

Based on the above discussions, in this paper, we propose a **LLM-Assisted Hyper-Relation Matching (LARM)** framework, including a LLM-assisted multimodal fusion module (LAM) and a hyper-relation matching module (HRM), to coherently address the question ❶-❷ in GFS-3Dseg. **To comprehensively enrich novel category prototypes**, LAM utilizes knowledge from both textual features and base classes visual features to compensate for the scarcity of novel class data. To utilize textual features, we first prompt LLM to generate diverse category-specific descriptions, then we propose a multimodal information distillation training strategy to implicitly embed textual information into the feature extractor, and a multimodal prototype fusion module to explicitly integrate textual knowledge into novel category prototypes. To utilize base class features, given that novel classes may likely share elemental patterns (meta-feature) with base classes [5, 36], we design a text-guided meta-feature selector to absorb novel-relevant information from base classes. Specifically, novel text features serve as a bridge to establish base-to-novel relationships along the channel dimension, guiding the selection of novel-relevant meta-features from base classes. These selected meta-features are then utilized to modulate novel class prototypes, further enhancing their representations. **To facilitate more accurate matching**, we conduct the hyper-relation matching based on the ranking distribution of query points with respect to the multimodal category prototypes (Figure 1). By integrating these strategies, LARM effectively tackles the two main challenges in GFS-3Dseg, leading to substantially improved performance.

Our contributions can be concluded as follows: (1) We conduct an in-depth analysis of the core challenges of GFS-3Dseg, *i.e.*, incomprehensive category representations and suboptimal matching processes, and propose to coherently tackle them in a unified and synergistic **LLM-Assisted Hyper-Relation Matching (LARM)** framework. (2) We leverage the extensive knowledge of large language models (LLMs) to augment vanilla category prototypes, enhancing category representations. Additionally, we propose a hyper-relation matching strategy to mitigate mismatches between points and category prototypes by modeling inter-class structural relationships. (3) Extensive experiments on various GFS-3Dseg datasets under different (*i.e.*, 1/5 shot) settings demonstrate that our approach significantly outperforms previous state-of-the-art GFS-3Dseg methods.

2. Related Work

Few-shot 3D Point Cloud Segmentation. Few-shot 3D point cloud semantic segmentation (FS-3Dseg) classifies points in novel categories using only a sparse set of labeled examples. Notable progress has been made in FS-3Dseg, benefiting from advancements in deep neural networks [14–16, 18, 20, 22, 23, 40, 41, 43, 44]. Existing methods typ-

ically follow a prototypical learning paradigm, condensing masked support features into one [21, 29] or multiple [1, 46] prototypes. AttMPTI [46] introduces an attention-aware multi-prototype transductive inference framework, enabling novel category classification with limited annotations. ProtoNet [29] is adapted to 3D contexts to generate class prototypes from support sets, with predictions based on cosine similarity to query features. To address contextual disparities, [21] proposes query-guided prototype refinement, while COSeg [1] optimizes query-support correlations by refining relationships between query points and prototypes. However, current FS-3Dseg methods focus solely on novel class segmentation, ignoring base classes encountered during training.

Generalized Few-shot 3D Point Cloud Segmentation. Generalized Few-Shot 3D Point Cloud Semantic Segmentation (GFS-3Dseg) has emerged as a promising paradigm to concurrently segment both base and novel classes within a unified framework during the inference stage. Xu *et al.*[38] pioneered the field of GFS-3Dseg by leveraging low-level geometric features of point clouds to complement and enhance novel class prototypes. Subsequently, Tsai *et al.*[31] propose utilizing background contextual information to improve the training of base class prototypes and the performance of few-shot learning. Although these efforts have demonstrated some success, their segmentation performance remains suboptimal due to the scarcity of support data and naive matching strategy. In contrast, our method enhances category representation by seamlessly integrating multimodal information via Language-Language Models (LLMs) and evolves the matching process through a well-designed hyper-relation matching strategy, surpassing the limitations of existing GFS-3Dseg methods.

3. Method

3.1. Problem Definition

Revisit the classic setting. The classic few-shot 3D point cloud segmentation follows the episodic paradigm [32], constructing disjoint training ($\mathcal{D}_{\text{base}}$) and testing ($\mathcal{D}_{\text{novel}}$) sets ($\mathcal{C}^b \cap \mathcal{C}^n = \emptyset$). Each episode, consisting of a support set S and a query set Q , trains the model to predict Q based on class information from S . The model learns on base classes \mathcal{C}^b and is evaluated on novel classes \mathcal{C}^n .

In generalized few-shot point cloud semantic segmentation (GFS-3Dseg), the dataset is divided into $\mathcal{D}_{\text{base}}$ with abundant labeled data and $\mathcal{D}_{\text{novel}}$ with K labeled samples per novel class, where $\mathcal{C}^b \cap \mathcal{C}^n = \emptyset$. The model is first trained on $\mathcal{D}_{\text{base}} = \{(\mathbf{P}_k^b, \mathbf{M}_k^b)\}_{k=1}^{|\mathcal{D}_{\text{base}}|}$, where $\mathbf{P}_k^b \in \mathbb{R}^{l \times d}$ is a point cloud and \mathbf{M}_k^b is the mask for base classes \mathcal{C}^b . Then, the model is fine-tuned using the limited novel data $\mathcal{D}_{\text{novel}} = \{(\mathbf{P}_k^{n,i}, \mathbf{M}_k^{n,i})_{i=1}^K\}_{k=1}^{|\mathcal{C}^n|}$. The testing dataset

$\mathcal{D}_{\text{test}} = \{(\mathbf{P}_k^q, \mathbf{M}_k^q)\}_{k=1}^{|\mathcal{D}_{\text{test}}|}$ contains query point clouds \mathbf{P}_k^q with ground-truth masks \mathbf{M}_k^q for both \mathcal{C}^b and \mathcal{C}^n . The objective of GFS-3Dseg is to segment both base and novel classes in the testing query point clouds.

3.2. Overview

As shown in Figure 2, the proposed LARM framework uses a two-phase training scheme: multimodal information distillation in the base-training phase, followed by a text-guided meta-feature selector, multimodal prototype fusion, and hyper-relation matching in the novel fine-tuning phase to collaboratively address challenges in GFS-3Dseg.

3.3. LLM-Assisted Multimodal Fusion

To improve category representation, especially for novel classes, LARM introduces several modules to fuse textual and visual modalities, aiming to enhance semantic richness and generalization, as described in detail below.

Multimodal information distillation training. To better utilize textual information to enhance multimodal information fusion, we employ contrastive learning for multimodal information distillation and alignment, which is jointly optimized with the training on base classes. Furthermore, to guarantee that the textual descriptions adequately capture the essential characteristics of point clouds, we adopt the 3D-specific heuristic prompt templates in [34, 50]:

Caption Generation: “Describe a point cloud of a [CLASS] in one sentence.”

Question Answering: “How to describe a point cloud of a [CLASS]?”

Paraphrase Generation: “Generate a synonym: A point cloud of a [CLASS].”

Words to Sentence: “Make a sentence with words: point cloud, [CLASS], obscure.”

Given an N -category point cloud dataset, we replace the “[CLASS]” placeholder in the prompt templates with each category name and input them to GPT-3 [3] to generate M descriptions per category, detailed category descriptions generated by GPT can be found in the Appendix C. These descriptions are encoded using CLIP’s textual encoder [28] and a text adapter to obtain their text features $\mathbb{T} = \{T_i\}_{i=1}^N$, where each $T_i = \{\mathbf{t}_{i,j}\}_{j=1}^M$ contains the text features of M descriptions for the i -th category. For a given N -category point cloud dataset, we define the contrastive loss as follows:

$$\mathcal{L}_{\text{con}} = -\frac{1}{N} \sum_{i=1}^N \left[\frac{1}{M} \sum_{j=1}^M \log \frac{\exp(\text{sim}(\mathbf{p}_i, \mathbf{t}_{i,j})/\tau)}{\sum_{k=1}^N \sum_{l=1}^M \exp(\text{sim}(\mathbf{p}_i, \mathbf{t}_{k,l})/\tau)} \right], \quad (1)$$

where \mathbf{p}_i represents the visual prototype of the i -th category obtained from the visual backbone. $\text{sim}(\mathbf{p}_i, \mathbf{t}_{i,j})$ is the similarity measure, typically defined as the cosine similarity:

$$\text{sim}(\mathbf{p}, \mathbf{t}) = \frac{\mathbf{p} \cdot \mathbf{t}}{\|\mathbf{p}\| \|\mathbf{t}\|}, \quad (2)$$

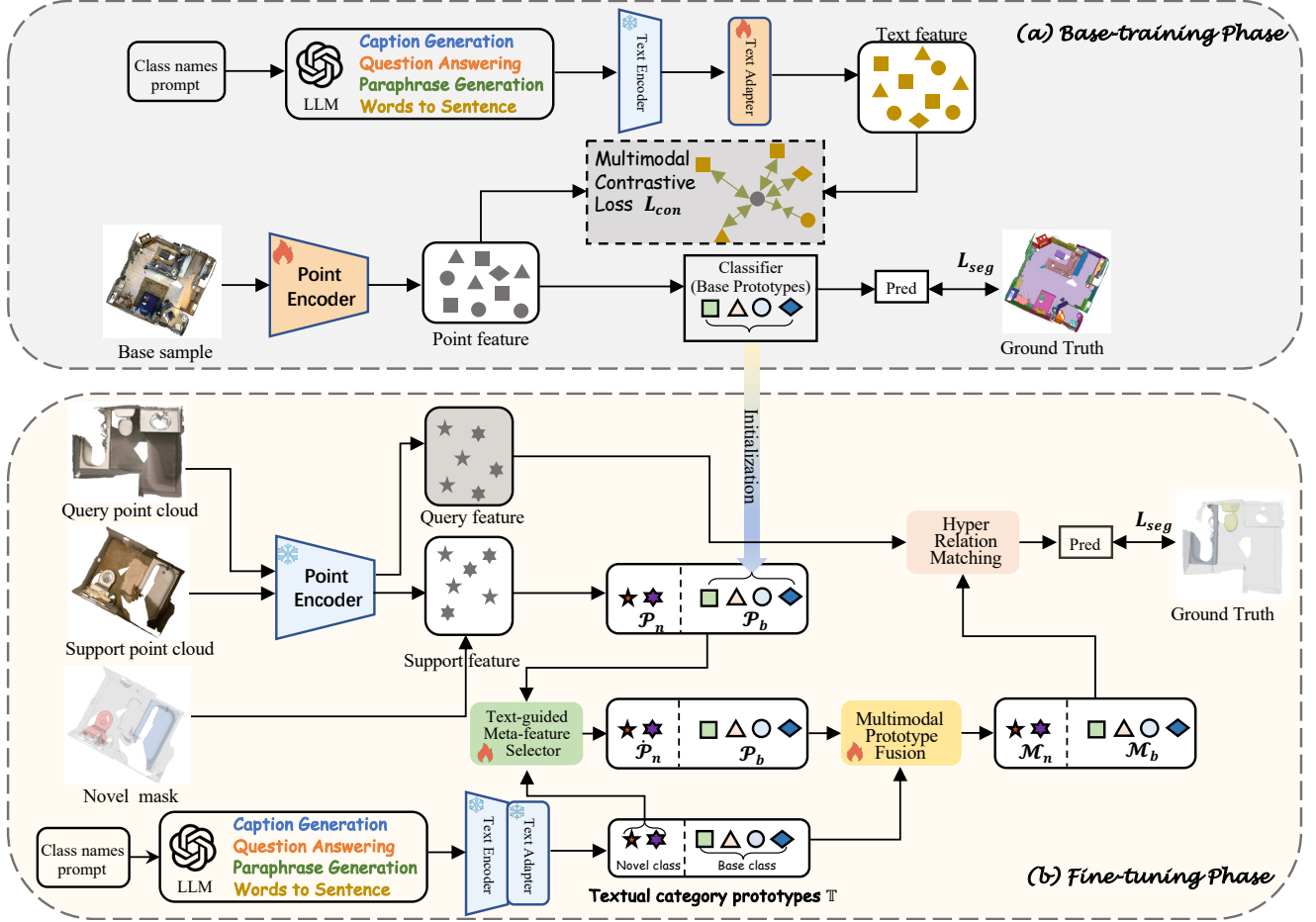


Figure 2. The overview of our proposed LLM-assisted hyper-relation matching (LARM) framework. (a) The base training phase jointly trains the point encoder, text adapter, and base prototypes. (b) In fine-tuning phase, the proposed text-guided meta-feature selector (TMS) and multimodal prototype fusion (MPF) are trained, and the base classes prototypes are directly inherited from the base training process as initialization. In the testing phase, the well-trained LARM is used to segment both base and novel classes in the query point cloud.

where $\tau > 0$ is a temperature parameter controlling the sharpness of the distribution. This function enforces the backbone to capture semantic consistency with textual information while maintaining discriminative power across categories, endowing it with multi-modal representation capabilities. More training details can be found [Appendix A](#).

Text-guided meta-feature selection. Through implicit distillation training, our backbone is endowed with multi-modal representation capabilities, enabling it to generate enhanced category prototypes \mathcal{P} . Since novel classes often share fundamental patterns with base classes [5, 36], we design a text-guided meta-feature selector to leverage helpful information from base classes to enhance novel category representation. In detail, for a given novel category n_i , we use its corresponding textual feature $T_{n_i} \in \mathbb{R}^{M \times d}$ as the query and base category prototypes as the keys and values to perform channel-wise attention, selecting features relevant to the novel category. Since each novel category has M textual descriptions from different perspectives, we

map a given base category prototype $\mathcal{P}_{b_i} \in \mathbb{R}^d$ into M sub-prototypes using a linear layer, yielding $\mathcal{P}_{b_i}^{sub} \in \mathbb{R}^{M \times d}$:

$$\mathcal{P}_{b_i}^{sub} = \mathcal{F}_{linear}(\mathcal{P}_{b_i}). \quad (3)$$

Then the input to our selector is designed as:

$$Q_1 = T_{n_i}^\top \mathbf{W}_{q1}, \quad K_1 = \mathcal{P}_{b_i}^\top \mathbf{W}_{k1}, \quad V_1 = \mathcal{P}_{b_i} \mathbf{W}_{v1}, \quad (4)$$

where $\mathbf{W}_{q1}, \mathbf{W}_{k1} \in \mathbb{R}^{M \times M'}$ and $\mathbf{W}_{v1} \in \mathbb{R}^{d \times d}$ are learnable parameters of the fully connected layers, which project the features and prototypes into their respective latent spaces. Then, we perform channel attention and obtain the channel-modulated base prototype:

$$\dot{\mathcal{P}}_{b_i} = \mathbf{W}_p((\text{softmax}(Q_1 \cdot K_1^\top)) \cdot V_1^\top)^\top, \quad (5)$$

where $\mathbf{W}_p \in \mathbb{R}^{d \times d}$ is the parameter of a fully connected layer. This channel selector allows the attention mechanism to compare the relevance between base prototypes and the

novel category from M different perspectives at the channel level and maintain the useful meta-feature to enhance novel prototypes. Now, we can employ the cross-attention to compute a weighted sum of the channel-modulated base prototypes $\dot{\mathcal{P}}_b$ for each novel prototype $\dot{\mathcal{P}}_{n_i}$. In specific:

$$\begin{aligned} \dot{\mathcal{P}}_{n_i} &= \mathcal{P}_{n_i} + \text{softmax}(Q_2 \cdot K_2^\top) \cdot V_2, \\ Q_2, K_2, V_2 &= \mathcal{P}_{n_i} W_{q2}, \dot{\mathcal{P}}_b W_{k2}, \dot{\mathcal{P}}_b W_{v2} \end{aligned} \quad (6)$$

where $W_{q2}, W_{k2} \in \mathbb{R}^{d \times d'}$, $W_{v2} \in \mathbb{R}^{d \times d}$ are fully connected layers. The generated $\dot{\mathcal{P}}_{n_i}$ absorb transferable knowledge from the rich base prototypes, further enriching the representational capacity of the novel prototypes. In Appendix B, we explain the rationale for using novel texts as queries to select information from base prototypes.

Multimodal prototype fusion. To explicitly incorporate text knowledge into visual category prototypes, we propose a multimodal prototype fusion method to integrate textual features \mathbb{T} into \mathcal{P}_b and $\dot{\mathcal{P}}_{n_i}$, resulting in multimodal prototypes using a shared attention module, formally defined as:

$$\begin{aligned} \mathcal{M}_{n_i} &= \dot{\mathcal{P}}_{n_i} + \text{softmax}(Q_3 \cdot K_3^\top) \cdot V_3, \\ Q_3, K_3, V_3 &= \dot{\mathcal{P}}_{n_i} W_{q3}, T_{n_i} W_{k3}, T_{n_i} W_{v3} \end{aligned} \quad (7)$$

$$\begin{aligned} \mathcal{M}_{b_i} &= \mathcal{P}_{b_i} + \text{softmax}(Q_3 \cdot K_3^\top) \cdot V_3, \\ Q_3, K_3, V_3 &= \mathcal{P}_{b_i} W_{q3}, T_{b_i} W_{k3}, T_{b_i} W_{v3} \end{aligned} \quad (8)$$

where $W_{q3}, W_{k3} \in \mathbb{R}^{d \times d'}$, $W_{v3} \in \mathbb{R}^{d \times d}$ are shared fully connected layers across both base and novel prototypes. These multimodal prototypes explicitly incorporate textual knowledge, further enhancing their robustness and richness.

3.4. Hyper-Relation Matching

Conventional point-to-prototype matching methods make classification decisions by selecting the category prototype that demonstrates the highest similarity to a given point. However, this approach overlooks correlations and similarities among different categories, leading to confusion and misclassification when dealing with semantically close categories. Merely choosing the most similar prototype fails to address blurred boundaries between categories, as similar categories may have heavily overlapping features and tightly clustered prototypes in the feature space, amplifying the probability of misclassification. To further harness the structure information implicit in the inter-class relationship for more effective matching, we introduce hyper-relation matching strategy. Specifically, given a query point feature $\mathcal{F}^q \in \mathbb{R}^{1 \times d}$ and our multimodal prototypes $\mathcal{M} = \{\mathcal{M}_{b_i}, \mathcal{M}_{n_i}\}$, where N represents the total number of categories with b base classes and n novel classes ($b + n = N$), we first select the top- K most similar prototypes from \mathcal{M} as *hyper-relation agents* $\mathcal{A} \in \mathbb{R}^{K \times d}$. We derive the scores of

point-agent relation $r^q \in \mathbb{R}^{1 \times K}$ between \mathcal{F}^q and \mathcal{A} using cosine similarity:

$$r^q = \frac{\mathcal{F}^q \mathcal{A}^\top}{\|\mathcal{F}^q\|_2 \|\mathcal{A}\|_2}. \quad (9)$$

The key insight is to model the *permutation distribution* of these agents rather than relying on deterministic similarity rankings. In other words, instead of assuming that only the permutation from largest to smallest exists, we consider that every permutation of the classes has a certain probability. The probability of one permutation $\pi \in \Pi$ (where $|\Pi| = K!$) given r (omit the point index q for convenience) can be calculated through chain rule factorization:

$$P(\pi|r) = \prod_{k=1}^K \frac{r_{\pi(k)}}{\sum_{k'=k}^K r_{\pi(k')}} \quad (10)$$

among which the $\pi(k)$ denotes the k^{th} class index of this permutation. For example, consider a given point where the selected agents correspond to the categories ‘‘bed’’, ‘‘table’’, and ‘‘chair’’. One possible permutation of these three agents is $\pi = (\text{table}, \text{bed}, \text{chair})$. The probability of this permutation π can be computed based on the point-agent relation scores r as follows:

$$P(\pi|r) = \frac{r(\text{table})}{r(\text{bed}) + r(\text{table}) + r(\text{chair})} \cdot \frac{r(\text{bed})}{r(\text{bed}) + r(\text{chair})}. \quad (11)$$

By associating the probabilities of all $|\mathcal{P}|$ permutations, we transform the individual point-agent relation scores r^q into class ranking probability distributions $P(\pi \in \mathcal{P}|r^q) \in \mathbb{R}^{1 \times |\mathcal{P}|}$. Similarly, we calculate the agent-ranking probability distributions for each multimodal category prototypes based on the same category agents, resulting in $P(\pi \in \mathcal{P}|r^p) \in \mathbb{R}^{1 \times |\mathcal{P}|}$, where r^p is derived from Eq. 9. To determine the classification result, the distribution of the query point is compared to the distributions of all prototypes using cosine similarity. The final prediction is given by:

$$\text{pred} = \arg \max_{r=1,2,\dots,N} [\text{cosine}(P(\pi \in \mathcal{P}|r^q), P(\pi \in \mathcal{P}|r^p))]. \quad (12)$$

The hyper-relation matching provides three key advantages:

- 1) The permutation distribution captures inter-class relationships and prevents overfitting to individual prototypes;
- 2) The top- K agent selection dynamically filters out irrelevant classes, focusing on semantically related ones;
- 3) The distributional similarity metric handles intra-class variations through probabilistic soft matching.

4. Experiments

4.1. Dataset and Evaluation Metric

Datasets. We conduct evaluations on two datasets including S3DIS[2] and ScanNet[7]. In specific, S3DIS contains 272

Table 1. Results on S3DIS under 5-shot and 1-shot settings.

Method	5-shot				1-shot			
	mIoU-B	mIoU-N	mIoU-A	HM	mIoU-B	mIoU-N	mIoU-A	HM
Fully Supervised	76.51	58.69	68.29	66.42	76.51	58.69	68.29	66.42
attMPTI [46]	34.90	16.08	26.21	21.99	21.89	11.39	17.05	14.95
PIFS [4]	56.99	19.66	39.76	29.23	57.85	14.59	37.88	23.31
CAPL [30]	73.56	35.18	55.85	47.51	72.80	23.87	50.22	35.67
GW [38]	73.61	43.26	59.60	54.42	74.10	29.66	53.58	41.92
PE [31]	74.77	50.23	63.44	60.06	74.54	39.78	58.50	51.34
Ours	75.64	54.08	65.68	63.06	75.68	44.52	61.29	56.04

Table 2. Results on ScanNet under 5-shot and 1-shot settings.

Method	5-shot				1-shot			
	mIoU-B	mIoU-N	mIoU-A	HM	mIoU-B	mIoU-N	mIoU-A	HM
Fully Supervised	43.12	37.04	41.34	39.85	43.12	37.04	41.34	39.85
attMPTI [46]	16.31	3.12	12.35	5.21	12.97	1.62	9.57	2.88
PIFS [4]	35.14	3.21	25.56	5.88	35.80	2.54	25.82	4.75
CAPL [30]	38.22	14.39	31.07	20.88	38.70	10.59	30.27	16.53
GW [38]	40.18	18.58	33.70	25.39	40.06	14.78	32.47	21.55
PE [31]	40.42	23.34	35.30	29.55	40.47	15.57	33.00	22.47
Ours	42.37	28.61	38.21	34.14	42.71	21.50	36.43	28.58

Table 3. Ablation study on S3DIS dataset under 1-shot setting. We evaluate the effectiveness of the multimodal information distillation training (MID), LLM-assisted multimodal fusion (LAM), which includes “text-guided meta-feature selection” (TMS) and “multimodal prototype fusion (MPF)”, and the hyper-relation matching (HRM).

MID	LAM		HRM	mIoU-B	mIoU-N	mIoU-A	HM
	TMS	MPF					
				74.10	29.66	53.58	41.92
✓				74.61	32.29	55.08	45.07
	✓			74.12	33.18	55.22	45.80
		✓		74.67	36.41	57.01	48.93
	✓	✓		74.65	39.34	58.35	51.52
✓	✓	✓		75.03	40.73	59.20	52.77
			✓	74.58	33.57	55.65	46.30
✓	✓	✓	✓	75.68	44.52	61.29	56.04

point clouds spanning six areas, annotated with 13 semantic classes. Area 6 is designated as the testing set (D_{test}), while the remaining five areas are used to create the training dataset, comprising both base and novel classes. ScanNet includes 1,513 point clouds annotated with 20 semantic classes. From this dataset, 1,201 point clouds are used to construct the training sets for D_{train}^b and D_{train}^n , while the remaining 312 point clouds form the test set D_{test} . For both datasets, the six classes with the fewest labeled points are selected as novel classes (C^n), while the rest are treated as base classes (C^b). This setup reflects a realistic scenario where novel classes are infrequent, and collecting sufficient training samples for them is challenging. As a result, the novel classes for S3DIS are table, window, column, beam, board, and sofa, while for ScanNet, they are sink, toilet, bathtub, shower curtain, picture, and counter. Following the

data pre-processing approach of [38], we divide each point cloud into $1\text{ m} \times 1\text{ m}$ blocks along the xy plane. From each block, we randomly sample $m = 2,048$ points to serve as input. Each input point has a feature dimension of $d_0 = 9$, including XYZ coordinates, RGB values, and normalized XYZ coordinates relative to the block.

Evaluation Metrics. We assess the model’s performance using the mean intersection-over-union (mIoU) metric. Specifically, we denote the mIoU for base classes, novel classes, and all classes as mIoU-B, mIoU-N, and mIoU-A, respectively. To provide a more balanced evaluation of the performance across both base and novel classes, we also compute the harmonic mean of mIoU-B and mIoU-N, *i.e.*, $\text{HM} = \frac{2 \times \text{mIoU-B} \times \text{mIoU-N}}{\text{mIoU-B} + \text{mIoU-N}}$. Unlike mIoU-A, the harmonic mean (HM) provides a fairer representation of the model’s performance across the two class types, as it does not overly favor the base classes [39].

4.2. Implementation details

Our approach is implemented in PyTorch [24], and all experiments are conducted on a single NVIDIA RTX 3090 GPU. We use GW [38] with a DGCNN [33] backbone pre-trained on base classes as the baseline. In the base-training phase, we integrate the proposed multimodal distillation into the training process, enhancing the backbone’s multimodal representation capabilities. The model is trained with a batch size of 32 for 150 epochs using the Adam optimizer, starting with a learning rate of 0.01, decayed by 0.5 every 50 epochs. The first three EdgeConv layers use pretrained weights with a learning rate of 0.001. During the novel

Table 4. Comparison of online computational costs and mIoU results on novel classes of the S3DIS dataset between our method and previous approaches under 1-shot settings.

Methods	#Params	FLOPs	FPS	Inference Time (ms)	1-shot
attMPTI [46]	357.82K	7.78G	1.58	632.91	11.39
GW [38]	355.08K	8.41G	39.52	25.30	29.66
Ours	1.73M	11.79G	23.21	43.08	44.52

Table 5. Offline processing time for category description generation and text feature extraction on S3DIS and ScanNet.

Phase	S3DIS	ScanNet
Description Generation: gpt-3.5-turbo	439.28 s	824.08 s
Text Feature Extraction: CLIP ViT-B/32	5.46 s	10.42 s
Total time	444.74 s	834.50 s

class fine-tuning phase, we randomly sample $K \in \{1, 5\}$ point clouds from novel classes as support samples. The model is trained with a batch size of 16 for 100 epochs using the Adam optimizer with an initial learning rate of 0.01. We use the “gpt-3.5-turbo” engine with a temperature of 0.7 to generate 24 unique descriptions per category.

4.3. Comparison with State-of-the-Art Methods

S3DIS. In Table 1, we present a comparison between our proposed method, LARM, and the state-of-the-art generalized few-shot point cloud segmentation approaches. The results clearly demonstrate that LARM achieves superior performance across all evaluated settings. Specifically, our method attains mIoU-N scores of 54.08% and 44.52% in the 5-shot and 1-shot scenarios, respectively, outperforming the strongest competitor, PE [31], by margins of 3.85% and 4.74%. Notably, the performance gains are even more significant in the 1-shot setting, where the limited support information for novel classes exacerbates the challenge of insufficient novel class representations. By leveraging multimodal fusion, our approach effectively incorporates textual knowledge as a complementary source, addressing this limitation and maintaining robust performance.

ScanNet. In Table 2, we present the comparison of results on the more challenging ScanNet dataset, which is characterized by its diverse and complex room types. Our proposed LARM continues to outperform the best existing method, achieving improvements of 5.27% and 5.93% mIoU in novel classes of the 5-shot and 1-shot settings, respectively. These results highlight the robustness and adaptability of our approach, especially in handling datasets with greater structural and semantic diversity.

Qualitative Results. Figure 3 presents our visualization results, demonstrating that our method achieves superior segmentation performance on novel categories (marked with red boxes.) while maintaining overall leading performance across both base and novel categories.

Computational Complexity. The proposed LARM extracts text features for all classes in the dataset using CLIP after generating class descriptions with LLMs in an offline

manner. Table 5 shows the offline time required for category description generation and text feature extraction on the S3DIS and ScanNet datasets. The offline time cost is minimal, while it effectively enables the utilization of rich semantic information from textual descriptions to enhance multimodal feature alignment, ensuring that the online inference remains efficient and unaffected. The precomputed text features are stored and directly loaded during training and testing, avoiding online regeneration and reducing computational costs. In Table 4, we compare the online costs of several methods. Our approach achieves better performance than previous methods with minimal increase in parameter count and computational overhead.

4.4. Ablation Study

We conduct extensive ablation studies to evaluate each component of LARM, with the first row of Table 3 showing results for the ablation baseline [38].

Effectiveness of module MID. The comparison between the 1st and 2nd rows of Table 3 reveals a significant performance improvement, with a 2.63% increase in mIoU-N and a 3.15% boost in HM. This highlights that the introduction of MID allows the feature extractor to distill knowledge from textual information, thereby enhancing its ability to produce more robust category representations.

Effectiveness of module LAM. Rows 3–6 of Table 3 present ablation studies on LAM’s core components, TMS and MPF, to verify their effectiveness individually.

1) *Effectiveness of TMS.* Comparing the 1st and 3rd rows in Table 3 shows that TMS improves the mIoU of novel categories by 3.52%, demonstrating its effectiveness. This is achieved through text-guided base meta-feature selection, allowing novel categories to leverage shared foundational patterns while reducing noise from base categories.

2) *Effectiveness of MPF.* As shown by the comparison between the 1st and 4th rows in Table 3, MPF improves the performance of both base and novel categories, with gains of 0.57% and 6.75%, respectively. This can be attributed to MPF leveraging the rich high-level semantic information from LLM-based text to enhance the original prototypes of both base and novel categories, thereby improving the robustness of the original visual features and boosting segmentation performance. Moreover, when the MID and TMS modules are incorporated alongside MPF, the model’s performance continues to improve. With all three modules combined, the performance increases by 0.93% for base categories and 11.07% for novel categories.

Effectiveness of module HRM. The comparison between the 7th and 1st rows in Table 3 shows that applying our HRM module alone improves performance by 0.48% for base categories and 3.91% for novel categories, thanks to its ability to leverage structural information to reduce intra-class variations and inter-class confusion. Furthermore,

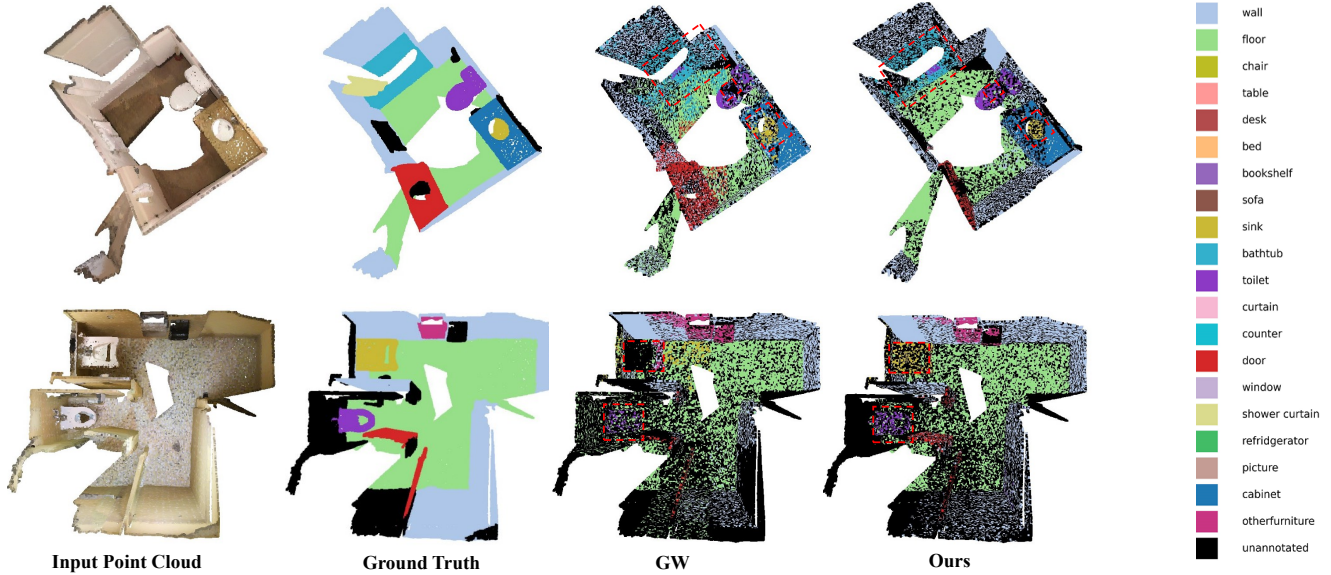


Figure 3. Qualitative comparison on 5-shot setting of ScanNet dataset. Note that some improvements are marked with red boxes.

Table 6. Per-class IoU results for the 5-shot setting on S3DIS, with new classes in red.

Method	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookcase	board	clutter
CAPL [30]	93.17	97.36	72.92	52.44	28.82	35.12	77.57	60.60	61.29	10.38	54.42	23.75	58.19
GW [38]	92.34	97.16	70.73	60.37	32.16	46.30	76.21	64.41	63.42	16.99	54.77	39.30	60.66
PE [31]	92.62	97.18	74.63	63.95	27.34	58.56	77.24	69.87	66.82	22.61	57.10	59.04	57.82
Ours	92.70	97.30	75.87	64.60	41.84	63.51	77.54	73.01	66.93	29.68	55.32	51.83	63.81

Table 7. Standard deviation.

Method	mIoU-B	mIoU-N	mIoU-A	HM
CAPL [30]	0.17	3.25	1.46	2.94
GW [38]	0.35	3.34	1.63	2.75
PE [31]	0.17	2.76	1.32	2.00
Ours	0.14	2.29	1.10	1.73

Table 8. Ablation study on the number of descriptions in S3DIS.

Description number M	mIoU-B	mIoU-N	mIoU-A	HM
$M = 4$	75.38	43.74	60.77	55.35
$M = 8$	75.70	44.05	61.09	55.69
$M = 12$	75.53	44.17	61.05	55.74
$M = 24$	75.68	44.52	61.29	56.04
$M = 40$	75.44	44.15	61.00	55.70

as shown in the 8th row, integrating all proposed modules achieves peak performance, with gains of 1.58% and 14.86% for base and novel categories, respectively, setting a new state-of-the-art for the GFS-3Dseg task and validating the effectiveness of our approach.

Per-class IoU and standard deviation. To provide a more detailed demonstration of the effectiveness of our method, Table 6 shows the per-class mIoU results. Our method achieves the best performance on most novel classes (5 out of 6) while maintaining leading performance on base classes. We also validate the robustness of LARM through standard deviation analysis, which is crucial for evaluating few-shot segmentation methods under diverse support conditions. Specifically, under the 5-shot setting of S3DIS, we sample 5 sets of support point clouds for each novel class, covering various geometric structures and color distributions. As shown in Table 7, our method achieves significantly lower standard deviation across all metrics, demonstrating robustness to support variability. This stability is

due to LARM’s ability to leverage LLMs for extracting high-level semantic features and addressing the scarcity of support data, while the hyper-relation matching mechanism ensures consistent query-support relationship modeling.

Impact of the number of generated descriptions. For each category, we utilize four different prompt formats, including *caption generation*, *question answering*, *paraphrase generation*, and *words to sentence*. In Table 8, we evaluate the effect of generating (1, 2, 3, 6, and 10 descriptions) for each prompt format, corresponding to a total of M (4, 8, 12, 24, and 40 descriptions). As shown, performance improves with an increasing number of descriptions and reaches its peak at $M=24$. However, further increasing the number of descriptions may result in redundant or repetitive textual information. Therefore, we adopt $M=24$ as the optimal setting in our experiments.

5. Conclusion

We propose a LLM-Assisted Hyper-Relation Matching (LARM) framework to tackle the challenges in generalized few-shot point cloud segmentation (GFS-3Dseg). LARM harnesses LLMs’ knowledge to enhance category prototypes and employs a hyper-relation matching strategy to assign points to categories by utilizing inter-class structural relationships. LARM surpasses state-of-the-art results and paves the way for further leveraging LLMs in GFS-3Dseg.

Acknowledgements

This work was supported by the National Defense Science and Technology Foundation Strengthening Program Funding (Grant 2023-JCJQ-JJ-0219).

References

- [1] Zhaochong An, Guolei Sun, Yun Liu, Fayao Liu, Zongwei Wu, Dan Wang, Luc Van Gool, and Serge Belongie. Rethinking few-shot 3d point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3996–4006, 2024. 1, 3
- [2] 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. 5
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2, 3
- [4] Fabio Cermelli, Massimiliano Mancini, Yongqin Xian, Zeynep Akata, and Barbara Caputo. Prototype-based incremental few-shot semantic segmentation. *arXiv preprint arXiv:2012.01415*, 2020. 6
- [5] Xinyue Chen, Miaoqing Shi, Zijian Zhou, Lianghua He, and Sophia Tsoka. Enhancing generalized few-shot semantic segmentation via effective knowledge transfer. *arXiv preprint arXiv:2412.15835*, 2024. 1, 2, 4
- [6] Yujia Chen, Rui Sun, Wangkai Li, Huayu Mai, Naisong Luo, Yuwen Pan, and Tianzhu Zhang. Alleviate and mining: Rethinking unsupervised domain adaptation for mitochondria segmentation from pseudo-label perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2339–2347, 2025. 2
- [7] 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. 5
- [8] Chenxun Deng, Dafang Li, Lin Ji, Chengyang Zhang, Baican Li, Hongying Yan, Jiyuan Zheng, Lifeng Wang, and Junguo Zhang. Chatdiff: A chatgpt-based diffusion model for long-tailed classification. *Neural Networks*, 181:106794, 2025. 2
- [9] Harintaka Harintaka and Calvin Wijaya. Improved deep learning segmentation of outdoor point clouds with different sampling strategies and using intensities. *Open Geosciences*, 16(1):20220611, 2024. 1
- [10] Shuting He, Xudong Jiang, Wei Jiang, and Henghui Ding. Prototype adaption and projection for few-and zero-shot 3d point cloud semantic segmentation. *IEEE Transactions on Image Processing*, 32:3199–3211, 2023. 1
- [11] Shuting He, Henghui Ding, Xudong Jiang, and Bihan Wen. Segpoint: Segment any point cloud via large language model. In *European Conference on Computer Vision*, pages 349–367. Springer, 2024. 1
- [12] Sheng Jin, Xueying Jiang, Jiaying Huang, Lewei Lu, and Shijian Lu. Llms meet vlms: Boost open vocabulary object detection with fine-grained descriptors. *arXiv preprint arXiv:2402.04630*, 2024. 2
- [13] Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20943–20953, 2024. 1
- [14] Wangkai Li, Rui Sun, Bohao Liao, Zhaoyang Li, and Tianzhu Zhang. Balanced learning for domain adaptive semantic segmentation. In *Forty-second International Conference on Machine Learning*. 2
- [15] Yanjun Li, Zhaoyang Li, Honghui Chen, and Lizhi Xu. Unbiased video scene graph generation via visual and semantic dual debiasing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19047–19056, 2025.
- [16] Zhaoyang Li, Wangkai Li, Huayu Mai, Tianzhu Zhang, and Zhiwei Xiong. Enhancing cell detection in histopathology images: a vit-based u-net approach. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 150–160. Springer, 2023. 2
- [17] Zhaoyang Li, Yuan Wang, Wangkai Li, Rui Sun, and Tianzhu Zhang. Localization and expansion: A decoupled framework for point cloud few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 18–34. Springer, 2024. 1
- [18] Zhaoyang Li, Yuan Wang, Wangkai Li, Tianzhu Zhang, and Xiang Liu. Dual-agent optimization framework for cross-domain few-shot segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9849–9859, 2025. 2
- [19] Jie Liu, Wenzhe Yin, Haochen Wang, Yunlu Chen, Jan-Jakob Sonke, and Efstratios Gavves. Dynamic prototype adaptation with distillation for few-shot point cloud segmentation. In *2024 International Conference on 3D Vision (3DV)*, pages 810–819. IEEE, 2024. 1
- [20] 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*, pages 3391–3401, 2024. 2
- [21] 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. 1, 3
- [22] 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. 2
- [23] Yuwen Pan, Rui Sun, Naisong Luo, Tianzhu Zhang, and Yongdong Zhang. Exploring reliable matching with phase

- enhancement for night-time semantic segmentation. In *European Conference on Computer Vision*, pages 408–424. Springer, 2024. 2
- [24] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6
- [25] 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. 1
- [26] 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. 1
- [27] Haoxuan Qu, Yujun Cai, and Jun Liu. Llms are good action recognizers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18395–18406, 2024. 2
- [28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3
- [29] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 1, 3
- [30] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11563–11572, 2022. 6, 8
- [31] Chih-Jung Tsai, Hwann-Tzong Chen, and Tyng-Luh Liu. Pseudo-embedding for generalized few-shot 3d segmentation. In *European Conference on Computer Vision*, pages 383–400. Springer, 2024. 1, 2, 3, 6, 7, 8
- [32] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 3
- [33] 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. 6
- [34] Lili Wei, Congyan Lang, Ziyi Chen, Tao Wang, Yidong Li, and Jun Liu. Generated and pseudo content guided prototype refinement for few-shot point cloud segmentation. *Advances in Neural Information Processing Systems*, 37:31103–31123, 2025. 1, 3
- [35] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4840–4851, 2024. 1
- [36] Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. Learning meta-class memory for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 517–526, 2021. 2, 4
- [37] Guoxin Xiong, Yuan Wang, Zhaoyang Li, Wenfei Yang, Tianzhu Zhang, Xu Zhou, Shifeng Zhang, and Zhang Yongdong. Aggregation and purification: Dual enhancement network for point cloud few-shot segmentation. In *IJCAI*, 2024. 1
- [38] Yating Xu, Conghui Hu, Na Zhao, and Gim Hee Lee. Generalized few-shot point cloud segmentation via geometric words. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21506–21515, 2023. 1, 2, 3, 6, 7, 8
- [39] Han-Jia Ye, Hexiang Hu, and De-Chuan Zhan. Learning adaptive classifiers synthesis for generalized few-shot learning. *International Journal of Computer Vision*, 129:1930–1953, 2021. 6
- [40] Chunhui Zhang, Guanjie Huang, Li Liu, Shan Huang, Yinan Yang, Xiang Wan, Shiming Ge, and Dacheng Tao. Webuav-3m: A benchmark for unveiling the power of million-scale deep uav tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9186–9205, 2022. 2
- [41] Chunhui Zhang, Xin Sun, Yiqian Yang, Li Liu, Qiong Liu, Xi Zhou, and Yanfeng Wang. All in one: Exploring unified vision-language tracking with multi-modal alignment. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5552–5561, 2023. 2
- [42] Canyu Zhang, Zhenyao Wu, Xinyi Wu, Ziyu Zhao, and Song Wang. Few-shot 3d point cloud semantic segmentation via stratified class-specific attention based transformer network. *arXiv preprint arXiv:2303.15654*, 2023. 1
- [43] Chunhui Zhang, Li Liu, Guanjie Huang, Hao Wen, Xi Zhou, and Yanfeng Wang. Webuot-1m: Advancing deep underwater object tracking with a million-scale benchmark. *Advances in Neural Information Processing Systems*, 37:50152–50167, 2024. 2
- [44] Chunhui Zhang, Li Liu, Hao Wen, Xi Zhou, and Yanfeng Wang. Mambatrack: Exploiting dual-enhancement for night uav tracking. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2
- [45] Nan Zhang, Zhiyi Pan, Thomas H Li, Wei Gao, and Ge Li. Improving graph representation for point cloud segmentation via attentive filtering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1244–1254, 2023. 1
- [46] 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. 1, 3, 6, 7
- [47] Qihao Zhao, Yalun Dai, Hao Li, Wei Hu, Fan Zhang, and Jun Liu. Ltgc: Long-tail recognition via leveraging llms-driven generated content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19510–19520, 2024. 2
- [48] Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Kun Kuang, and Chao Wu. Model tailor:

Mitigating catastrophic forgetting in multi-modal large language models. *arXiv preprint arXiv:2402.12048*, 2024.

- [49] Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. Llafs: When large language models meet few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3065–3075, 2024.
- [50] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Point-clip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2639–2650, 2023. [2](#), [3](#)