Rethinking Multimodal Few-Shot 3D Point Cloud Segmentation: From Fused Refinement to Decoupled Arbitration

Wentao Bian, Fenglei Xu*
School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou, China
bwt1819548635@gmail.com, xufl@mail.usts.edu.cn

Abstract

In this paper, we revisit multimodal few-shot 3D point cloud semantic segmentation (FS-PCS), identifying a conflict in "Fuse-then-Refine" paradigms: the "Plasticity-Stability Dilemma." In addition, CLIP's inter-class confusion can result in semantic blindness. To address these issues, we present the **D**ecoupled-experts **A**rbitration **F**ew-**S**hot **S**egNet (**DA-FSS**), a model that effectively distinguishes between semantic and geometric paths and mutually regularizes their gradients to achieve better generalization. DA-FSS employs the same backbone and pre-trained text encoder as MM-FSS to generate text embeddings, which can increase free modalities' utilization rate and better leverage each modality's information space. To achieve this, we propose a Parallel Expert Refinement module to generate each modal correlation. We also propose a **S**tacked **A**rbitration **M**odule (**SAM**) to perform convolutional fusion and arbitrate correlations for each modality pathway. The Parallel Experts decouple two paths: a Geometric Expert maintains plasticity, and a Semantic Expert ensures stability. They are coordinated via a **D**ecoupled **A**lignment **M**odule (**DAM**) that transfers knowledge without propagating confusion. Experiments on popular datasets (S3DIS, ScanNet) demonstrate the superiority of DA-FSS over MM-FSS. Meanwhile, geometric boundaries, completeness, and texture differentiation are all superior to the baseline. The code is available at: https://github.com/MoWenQAQ/DA-FSS/.

1 Introduction

3D point cloud analysis is following the development of **Label-efficient learning** [Xiao *et al.*, 2024; Chen *et al.*, 2023] and the advent of Large Vision-Language Models (VLMs) and 3D-LLMs [Hong *et al.*, 2023] has triggered a revolution in 3D scene understanding [Xu *et al.*, 2024; Xue *et al.*, 2024; Guo *et al.*, 2023; Hong *et al.*, 2023; Liang *et al.*, 2024]. Through establishing a connection between point clouds and human language, some methods have broken the recognition bottleneck imposed by the closed label sets of traditional supervised learning [Qi *et al.*, 2017a; Wang *et al.*, 2019; Wu *et al.*, 2024b]. Meanwhile, the paradigm shift towards leveraging foundation models is exemplified by Point-Bind [Guo *et al.*, 2023] and ULIP-2 [Xue *et al.*, 2024; Xue *et al.*, 2023] which are two examples of aggregated frameworks that have successfully created a **shared embedding space** for 3D and human language. For instance, OpenScene [Peng *et al.*, 2023], OV3D [Jiang *et al.*, 2024], RegionPLC [Yang *et al.*, 2024], UniM-OV3D [He *et al.*, 2024], Open3DIS [Nguyen *et al.*, 2024], and PLA [Ding *et al.*, 2023] all demonstrate solid zero-shot generalization by building upon these semantic priors. Despite these advances, Few-Shot Learning (FSL) [Snell *et al.*, 2017; Zhao *et al.*, 2021; Hong *et al.*, 2022] is still a critical challenge due to annotated scarce data.

In the realm of MM-FS-PCS, our key insight is that **MM-FSS** [An *et al.*, 2025b] is the pioneering **SOTA** that validated the utility of VLM priors. However, we've found that MM-FSS suffers from a *Plasticity-Stability Dilemma* which creates a **Gradient Domination**. This is caused by CLIP's category confusion [Li *et al.*, 2025; Zhu *et al.*, 2023; Tu *et al.*, 2023; Qiu *et al.*, 2024], which leads to *semantic hallucinations* in geometrically ambiguous regions (e.g., confusing a white wall with a picture).

Specifically, CLIP (Contrastive Language-Image Pre-training) produces semantic features that have been trained at a large scale. Its vector norms are typically large and stable. MM-FSS chooses to freeze the parameters of CLIP (IF Head) to preserve semantic features, leveraging the VLM, aligning with the trend of parameter-efficient fine-tuning [Tang *et al.*, 2024; Zhang *et al.*, 2022]. This leads to the problem that the model will discover that "I can achieve a high score simply by copying CLIP's semantic outputs." Thereby causing the 3D encoder to learn from 2D features in a simplistic manner, resulting in feature representations with relatively smaller norms, with semantic features dominating gradient flow.

Motivated by these important observations, a pertinent question arises: How can we limit gradient conflict and prevent category confusion? In this paper, the idea comes from the philosophy of Logits DeConfusion [Li *et al.*, 2025], which

*Corresponding author

physically separates confusion noise through residual learning. So we suggest a new paradigm **"Decoupled Arbitration"** to resolve this problem. We believe that geometric adaptation and semantic preservation need pathways that are completely separate from each other. Interaction should be limited to two laws: **Soft Regularization** which guide the geometric expert without messing up its input and **Late Arbitration** which only combines decisions after the decoupled decisions have been independently refined. Specifically, we introduce a novel model, **Decoupled-experts Arbitration Few-Shot SegNet (DA-FSS)**, to leverage multimodal information more efficiently. Concretely, we develop an **Adaptive Geometric Expert** and a **Static Semantic Expert** to form the Decoupled-experts. The following **S**tacked **A**rbitration **M**odule (**SAM**) performs convolutional fusion of the independently refined features from both pathways, building upon utilizing semantic guidance from textual information. Additionally, we propose **D**ecoupled **A**lignment **M**odule (**DAM**) enforcing the geometric expert to align with the semantic expert at the representation level so as to prevent the decoupling from causing each path to deviate excessively.

We systematically compare our DA-FSS against existing methods AttMPTI [Zhao *et al.*, 2021], QGE [Ning *et al.*, 2023], COSeg [An *et al.*, 2024], and MM-FSS [An *et al.*, 2025b] on S3DIS [Armeni *et al.*, 2016] and ScanNet [Dai *et al.*, 2017] datasets 4.2, suggesting the superiority of DA-FSS across various settings.

The contributions of our work are threefold: (i) We design a novel model, DA-FSS, to more effectively exploit information from different modalities. To the best of our knowledge, this is the first work to explore decoupling in MM-FSS. (ii) Learning from the baseline's novel cost-free multimodal FS-PCS design philosophy, our DA-FSS efficiently reuses the pre-trained alignment already completed by the baseline to achieve optimization without any extra parameter overhead. (iii) Extensive experimental results demonstrate the value of our proposed framework and its consistent efficacy across different few-shot settings. This work will provide a new perspective for other multimodal fusion approaches.

2 Related Work

2.1 Few-Shot 3D Point Cloud Segmentation

Although fully-supervised methods have achieved promising results in 3D point cloud segmentation [Qi *et al.*, 2017a; Qi *et al.*, 2017b; Wang *et al.*, 2019; Lai *et al.*, 2022; Wu *et al.*, 2024b; Yu *et al.*, 2022], the expensive annotation costs have shifted research attention toward few-shot learning. As a pioneer, AttMPTI [Zhao *et al.*, 2021] proposed a transductive label propagation method. Later works mainly concentrate on narrowing the semantic gap [Zhang *et al.*, 2023; Ning *et al.*, 2023; Ning *et al.*, 2023; Xiong *et al.*, 2024] and improving feature representation [An *et al.*, 2024]. Specifically, AttMPTI [Zhao *et al.*, 2021] and COSeg [An *et al.*, 2024] optimize prototype correlations to reduce intra-class variance. In contrast, SCAT [Zhang *et al.*, 2023] and QGE [Ning *et al.*, 2023] apply advanced cross-attention mechanisms for fine-grained alignment between support and query features. However, recent observations indicate that these methods rely mainly on geometric features, which limits their generalization in complex open-world scenes (e.g., distinguishing objects with similar geometry).

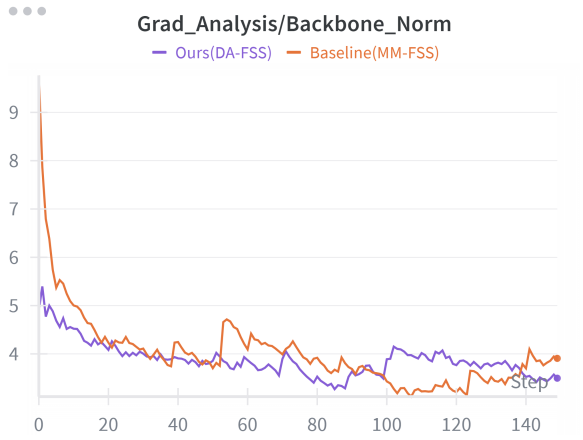

Figure 1: The Baseline (Orange)'s gradient norm declines rapidly with high volatility, characterized by diminishing gradients due to semantic domination. In contrast, our DA-FSS (Purple)'s gradient norm maintains a relatively stable trend.

2.2 Multimodal 3D Point Cloud Segmentation

With the rapid development of unimodal baselines, researchers are increasingly exploring multimodality for further performance gains. Language is currently the most popular auxiliary modality. Unified frameworks [Guo *et al.*, 2023; Xue *et al.*, 2024; Xu *et al.*, 2024; Zhou *et al.*, 2024] align 3D point clouds with image-text spaces to boost performance . Furthermore, open-vocabulary methods (e.g., OV3D [Jiang *et al.*, 2024], RegionPLC [Yang *et al.*, 2024], Open3DIS [Nguyen *et al.*, 2024], UniM-OV3D [He *et al.*, 2024], PointTFA [Wu *et al.*, 2024a]) utilize VLM guidance to achieve open-set segmentation. Nevertheless, these approaches are tailored primarily for fully supervised or zero-shot settings. In the context of FS-PCS, existing works still predominantly use unimodal input. Recently, MM-FSS [An *et al.*, 2025b] and Generalized FS-PCS [An *et al.*, 2025a] attempt to leverage multimodal information. Critically, these methods follow a "Fused Refinement" paradigm. We observe that the aggressive fusion in these works leads to *Gradient Domination*. To the best of our knowledge, we are the first to introduce decoupled expert arbitration [Li *et al.*, 2025; Zhang *et al.*, 2022; Zhao *et al.*, 2022] into multimodal FS-PCS.

3 Methodology

3.1 Problem Setup

Multimodal FS-PCS. Following prior works [An *et al.*, 2024; An *et al.*, 2025b], we formulate this task as the popular episodic paradigm. Each N-way K-shot episode comprises a support set $\mathcal{S} = \{X_s^{n,k}, Y_s^{n,k}\}$ and a query set $\mathcal{Q} = \{X_q, Y_q\}$, where X and Y denote the point cloud and corresponding labels, respectively. The goal is to segment query samples X_q into N target classes and 'background' by

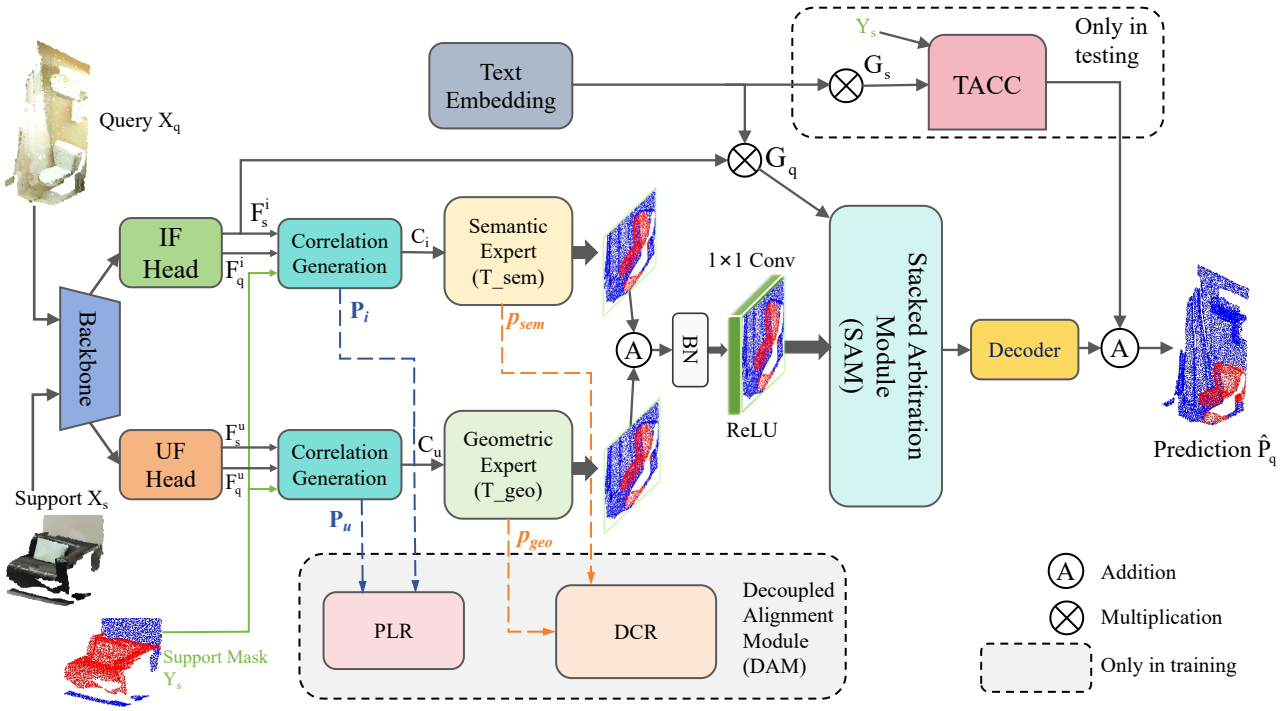

Figure 2: **Overall architecture of the proposed DA-FSS.** Given support and query point clouds, we first generate intermodal correlations $F^i_{s/q}$ from the IF head and unimodal correlations $F^u_{s/q}$ from the UF head. These correlations are then forwarded to the **Parallel Experts** (Geometric Expert $\mathcal{T}_{geo}$ and Semantic Expert $\mathcal{T}_{sem}$) to independently refine features, isolating plasticity from stability. Moreover, we use the **Decoupled Alignment Module (DAM)** to align correlations (C_i, C_u) via PLR and refined experts (P_{sem}, P_{geo}) via DCR **only in training**. Finally, we propose the **Stacked Arbitration Module (SAM)**, which synthesizes the final decision by leveraging boundary-injected guidance (G_{base} and G_q) to effectively arbitrate between geometric and semantic pathways after a 1×1 convolutional fusion. For clarity, we present the model under the 1-way 1-shot setting.

leveraging knowledge from $\mathcal{S}$. Following the multimodal setting, our DA-FSS introduces textual and 2D modalities.

3.2 Motivation

Current SOTA methods [An *et al.*, 2025b] typically adopt an early fusion strategy, obtaining a joint representation $H = \sigma(W \cdot [F_{geo} \oplus F_{sem}])$. Although this is a simple and intuitive approach, we argue that this paradigm creates a fundamental optimization conflict : during meta-training, the gradient flow to the geometric encoder called **UF Head** [An *et al.*, 2025b] is always coupled with the semantic branch called **IF Head** [An *et al.*, 2025b]:

$$\frac{\partial \mathcal{L}}{\partial F_{UF}} = \frac{\partial \mathcal{L}}{\partial H} \cdot \frac{\partial H}{\partial F_{UF}} \tag{1}$$

We identify two issues in this interaction: (1) The static semantic features F_{sem} (from frozen VLMs) typically exhibit significantly larger norms than the adaptive geometric features, numerically overwhelming the sparse geometric gradients. (2) Frozen VLMs suffer from inherent *texture bias* [Li *et al.*, 2025; Tu *et al.*, 2023]. In geometrically ambiguous regions, the optimization shortcut provided by texture semantics suppresses the learning of structural details.

Consequently, the **UF Head** fails to adapt effectively, a phenomenon we term *Plasticity Decay*:

$$\|\nabla_{\theta_{UF}} \mathcal{L}\| \xrightarrow{training} \delta_{low} \ll \|\nabla_{\text{init}}\| \tag{2}$$

We provide concrete evidence of this phenomenon in Figure 1. By monitoring the training dynamics, we observe that the baseline's geometric gradients (Orange) are strictly suppressed by the semantic branch, effectively rendering the backbone "lazy" so that the **UF Head** will only obey the VLM's guidance from the **IF Head** (**Frozen**) without learning to optimize itself. This observation motivates us to **physically decouple** the optimization pathways to restore the backbone's vitality.

3.3 Method Overview

The overall architecture of the proposed DA-FSS is depicted in Figure 2. Given support and query point clouds, we first generate two sets of dense correlations: intermodal correlations from the Intermodal Feature (IF) head and unimodal correlations from the Unimodal Feature (UF) head. Both intermodal and unimodal correlations are then forwarded to the **Parallel Expert Refinement** module to independently produce refined geometric and semantic features. Beyond independent refinement, we use the **D**ecoupled **A**lignment **M**odule (**DAM**) to bridge the modal gap. This module acts as a soft regularizer to align the experts without gradient interference. We then exploit useful boundary-injected guidance to synthesize the final decision in the **S**tacked **A**rbitration **M**odule (**SAM**).

Existing MultiModal FS-PCS approaches [An *et al.*, 2025b] typically have two training steps: a pretraining step

for obtaining an effective feature extractor, and a meta-learning step towards few-shot segmentation tasks. Our method follows this two-step training paradigm. First, we pretrain the backbone and IF head using 3D point clouds and 2D images. To more objectively demonstrate our improvement results, we chose to directly use the same method as MM-FSS does, or even download officially released pretrained weights directly. Second, we conduct meta-learning to train the model end-to-end while freezing the backbone and IF head to maintain stability. In the following, we elaborate on Parallel Experts, DAM, and SAM modules.

3.4 Parallel Experts Refinement

Inspired by Mixture-of-Experts (MoE) [Jacobs *et al.*, 1991], we propose the Parallel Experts module, as illustrated in Figure 2 to preserve the unique characteristics of each modality. Following [An *et al.*, 2025b], we first merge correlations to capture support-query interactions. The *Unimodal Correlation* $C^u \in \mathbb{R}^{N_q \times N_s}$ captures geometric similarity, while the *Intermodal Correlation* $C^i \in \mathbb{R}^{N_q \times N_s}$ captures semantic affinity using VLM projections. This enhanced comprehension promotes the transfer of knowledge from the support set to the query point cloud, thereby refining segmentation outcomes.MM-FSS chose to sum C^u and C^i, but we do not opt for this approach; conversely, we propose that:

Adaptive Geometric Expert ($\mathcal{T}_{geo}$). This expert is fully learnable and responsible for *Plasticity* by receiving C^u, which focuses on capturing task-specific structural variations. To this end, we employ a standard Transformer Layer (incorporating MHSA) to capture the holistic feature tensor of H_{geo} (192-dimensional):

$$\begin{aligned} H_{geo} &= \mathcal{F}_{lin1}(C^u) \\ R_{geo} &= \text{LN}\left(H_{geo} + \text{MHSA}(H_{geo}, H_{geo}, H_{geo})\right) \end{aligned} \tag{3}$$

Static Semantic Expert ($\mathcal{T}_{sem}$). This expert is fully learnable and responsible for *Stability* by receiving C^i, which leverages the frozen VLM priors. This is to provide a consistent semantic reference robust to few-shot noise. Although this expert's functions differ, we have chosen an identical design but H_{sem} (512-dimensional):

$$\begin{aligned} H_{sem} &= \mathcal{F}_{lin2}(C^i) \\ R_{sem} &= \text{LN}\left(H_{sem} + \text{MHSA}(H_{sem}, H_{sem}, H_{sem})\right) \end{aligned} \tag{4}$$

Crucially, by isolating these pathways ($\mathcal{I}(R_{geo}; R_{sem}) \approx 0$ during refinement), we ensure the experts are optimized solely based on features unique to their respective responsibilities, thereby protecting UF head against semantic gradient domination.

3.5 Decoupled Alignment Module (DAM)

While the Parallel Experts module effectively isolates interference between modalities, the consistency would largely disappear due to decoupling, which could lead to feature divergence. However, we cannot forcibly intervene to rigidly align the features of the two modules. Therefore, we propose the DAM, as illustrated in Fig. 2. DAM integrates structural consensus from the semantic expert to regularize the geometric expert. Additionally, since raw semantic priors contain fixed confusion patterns [Li *et al.*, 2025], DAM employs a stop-gradient strategy to filter out texture noise, accounting for the reliability gap between adaptive and frozen features.

Prototype Loss Regularization (PLR). Given learnable geometric prototypes P_u and frozen semantic prototypes P_i, since the semantic prototypes P_i originate from stable VLM priors, they provide informative guidance on the true semantic manifold. Therefore, we first compute the alignment distance between them to generate structural constraints $\mathcal{L}_{PLR}$. Specifically, we project P_u (192-dimensional) to the dimension of P_i (512-dimensional) via a linear layer F_{proj} and minimize:

$$\mathcal{L}_{PLR} = \frac{1}{|B_S|} \sum_{(P_u, P_i) \in B_S} \|F_{proj}(P_u) - \text{sg}(P_i)\|_2^2, \tag{5}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. Note that this effectively uses the semantic prototype as a "Teacher" anchor, ensuring the geometric encoder learns the semantic structure without inheriting confusion gradients.

Decoupled Consistency Regularization (DCR). Next, to enforce consistency at the decision boundary, we minimize the divergence between the probability distributions of the two experts (p_{geo}, p_{sem}). Since confusion patterns are often manifested in the logits distribution, the regularization term is computed as follows:

$$\begin{aligned} \mathcal{L}_{DCR} = &\frac{1}{2} D_{KL}(p_{geo} \parallel \text{sg}(p_{sem})) \\ &+ \frac{1}{2} D_{KL}(p_{sem} \parallel \text{sg}(p_{geo})). \end{aligned} \tag{6}$$

Through this interaction, $\mathcal{T}_{sem}$ can also learn the specialized adaptation of the other branch, rather than solely optimizing itself. This is akin to how a "teacher" typically influences a student, yet the "student" can also subtly impact their own "teacher."

This DAM module fully leverages the stability of the semantic expert to guide the plasticity of the geometric expert, helping to maintain feature coherence. Note that it uses stop-gradient operators to prevent the propagation of confusion noise, improving the effective collaboration of decoupled experts.

3.6 Stacked Arbitration Module (SAM)

While the two experts can make predictions independently; without fusing and refining these two predictions, the multimodal system cannot fully realize its potential for cross-checking and compensating for deficiencies. To synthesize the final prediction from the refined experts, we propose the Stacked Arbitration Module (SAM). Unlike prior works that rely on simple concatenation, we opt for a **Stacked Transformer Arbitration** architecture to model global interactions between geometric and semantic evidences.

Feature Merging Layer. We first project the concatenated expert outputs into a unified space (192-dimensional). Consistent with our implementation, Batch Normalization (BN) is applied pre-fusion to calibrate the scale discrepancy between adaptive and static features, followed by a 1×1 convolution and ReLU activation:

$$R_{merged} = \text{ReLU}(\mathcal{F}_{conv}(\text{BN}([R_{geo} \parallel R_{sem}]))) \tag{7}$$

Method	1-way 1-shot			1-way 5-shot			2-way 1-shot			2-way 5-shot		
	S^0	S^1	Mean	S^0	S^1	Mean	S^0	S^1	Mean	S^0	S^1	Mean
Official Results (from Literature)												
AttMPTI [Zhao *et al.*, 2021]	36.32	38.36	37.34	46.71	42.70	44.71	31.09	29.62	30.36	39.53	32.62	36.08
QGE [Ning *et al.*, 2023]	41.69	39.09	40.39	50.59	46.41	48.50	33.45	30.95	32.20	40.53	36.13	38.33
COSeg [An *et al.*, 2024]	47.17	48.37	47.77	50.93	49.88	50.41	37.15	38.99	38.07	42.73	40.25	41.49
MM-FSS [An *et al.*, 2025b]	49.84	54.33	52.09	51.95	56.46	54.21	41.98	46.61	44.30	46.02	54.29	50.16
Reproduced Environment (Strict Control)												
MM-FSS† (Official Weights)	48.50	54.15	51.33	51.76	56.43	54.10	41.70	47.00	44.35	45.67	54.82	50.25
Ours (DA-FSS)	**49.14**	**55.93**	**52.54**$_{\uparrow 1.21}$	**51.90**	**60.45**	**56.18**$_{\uparrow 2.08}$	**42.07**	**47.16**	**44.62**$_{\uparrow 0.27}$	**45.78**	**56.30**	**51.04**$_{\uparrow 0.79}$

Table 1: **Quantitative results on S3DIS.** *Top:* Official results reported in literature. *Bottom:* Comparison under our strictly controlled evaluation protocol using official pre-trained weights for the baseline (MM-FSS†). Our method (DA-FSS) consistently outperforms the baseline across most splits.

Method	1-way 1-shot			1-way 5-shot			2-way 1-shot			2-way 5-shot		
	S^0	S^1	Mean	S^0	S^1	Mean	S^0	S^1	Mean	S^0	S^1	Mean
Official Results (from Literature)												
AttMPTI [Zhao *et al.*, 2021]	34.03	30.97	32.50	39.09	37.15	38.12	25.99	23.88	24.94	30.41	27.35	28.88
QGE [Ning *et al.*, 2023]	37.38	33.02	35.20	45.08	41.89	43.49	26.85	25.17	26.01	28.35	31.49	29.92
COSeg [An *et al.*, 2024]	41.95	42.07	42.01	48.54	44.68	46.61	29.54	28.51	29.03	36.87	34.15	35.51
MM-FSS [An *et al.*, 2025b]	46.08	43.37	44.73	54.66	45.48	50.07	43.99	34.43	39.21	48.86	39.32	44.09
Reproduced Environment (Strict Control)												
MM-FSS† (Official Weights)	45.68	**43.23**	44.46	54.80	45.79	50.30	44.01	34.44	39.23	49.11	39.48	44.20
Ours (DA-FSS)	**48.95**	41.96	**45.46**$_{\uparrow 1.00}$	**55.40**	**46.30**	**50.85**$_{\uparrow 0.55}$	**45.01**	**34.53**	**39.77**$_{\uparrow 0.54}$	**50.61**	**40.45**	**45.53**$_{\uparrow 1.33}$

Table 2: **Quantitative results on ScanNet.** *Top:* Literature results. *Bottom:* Comparison under our strictly controlled evaluation protocol using official pre-trained weights. mIoU results are reported in %. The best performance in the controlled group is highlighted in **bold**.

Dual-Guidance Injection. We inject inductive biases strictly at the boundaries using a "Negative-Suppression, Positive-Enhancement" strategy:

- **Input: Background Suppression (G_{base}).** At the input of every Transformer layer, we explicitly inject Base-Class Guidance (G_{base}) to refine the background representation. Specifically, we concatenate G_{base} with the background token R_{bg} to suppress easy negatives early, while leaving foreground tokens R_{fg} untouched:

$$R_{bg}^{(1)} = \mathcal{F}_{lin}([R_{bg} \parallel G_{base}]), \quad R_{in}^{(1)} = [R_{bg}^{(1)} \parallel R_{fg}] \quad (8)$$

  This can lend the model a more hierarchical structure, enabling further refinement of the feature space.

- **Output: Semantic Gating (G_q).** At the final layer, we employ a multiplicative gating mechanism derived from the target class embedding (G_q). This acts as a semantic gate to enhance foreground precision by scaling the feature magnitude based on textual affinity:

$$R_{final} = R_{arb}^{(N)} \odot (1 + \sigma(G_q)) \quad (9)$$

  where $\odot$ denotes element-wise multiplication and σ is a scaling function.

This effectively mitigates issues of over-suppression or over-amplification while also reducing the parameter overhead in deep stacks beyond two layers.

3.7 Optimization Objective

After the Stacked Arbitration Module (SAM), the synthesized feature tensor R_{final} is transformed into the final prediction $\hat{P}_q \in \mathbb{R}^{N_q \times (N_{way}+1)}$ by a decoder comprising a KPConv [Thomas *et al.*, 2019] layer and an MLP classifier. The whole model is optimized end-to-end by minimizing a compound objective function. To ensure expert coordination without compromising individual plasticity, we integrate the segmentation task with decoupled alignment constraints:

$$\begin{aligned}\mathcal{L}_{total} =& \mathcal{L}_{seg} + \lambda_{base}\mathcal{L}_{base} \\ &+ \lambda_{PLR}\mathcal{L}_{PLR} + \lambda_{DCR}\mathcal{L}_{DCR},\end{aligned} \quad (10)$$

where $\mathcal{L}_{seg}$ is the standard cross-entropy loss between $\hat{P}_q$ and the ground-truth Y_q, and $\mathcal{L}_{base}$ checks the predictions for the base class. This formulation ensures a cooperative optimization: vertical gradients from $\mathcal{L}_{seg}$ optimize the decision boundary, while horizontal gradients from the alignment terms maintain expert coordination.

4 Experiments

In this section, we conduct extensive experiments to validate our proposed Decoupled-experts Arbitration (DA-FSS) framework. We first introduce the experimental setup. Then, we compare our method with state-of-the-art (SOTA) approaches on two challenging benchmarks: S3DIS [Armeni *et al.*, 2016] and ScanNet [Dai *et al.*, 2017]. Finally, we perform detailed ablation studies to demonstrate the effectiveness of each key component in our architecture.

4.1 Experimental Setup

Datasets. We evaluate our method on two standard few-shot 3D point cloud segmentation benchmarks: S3DIS [Ar-

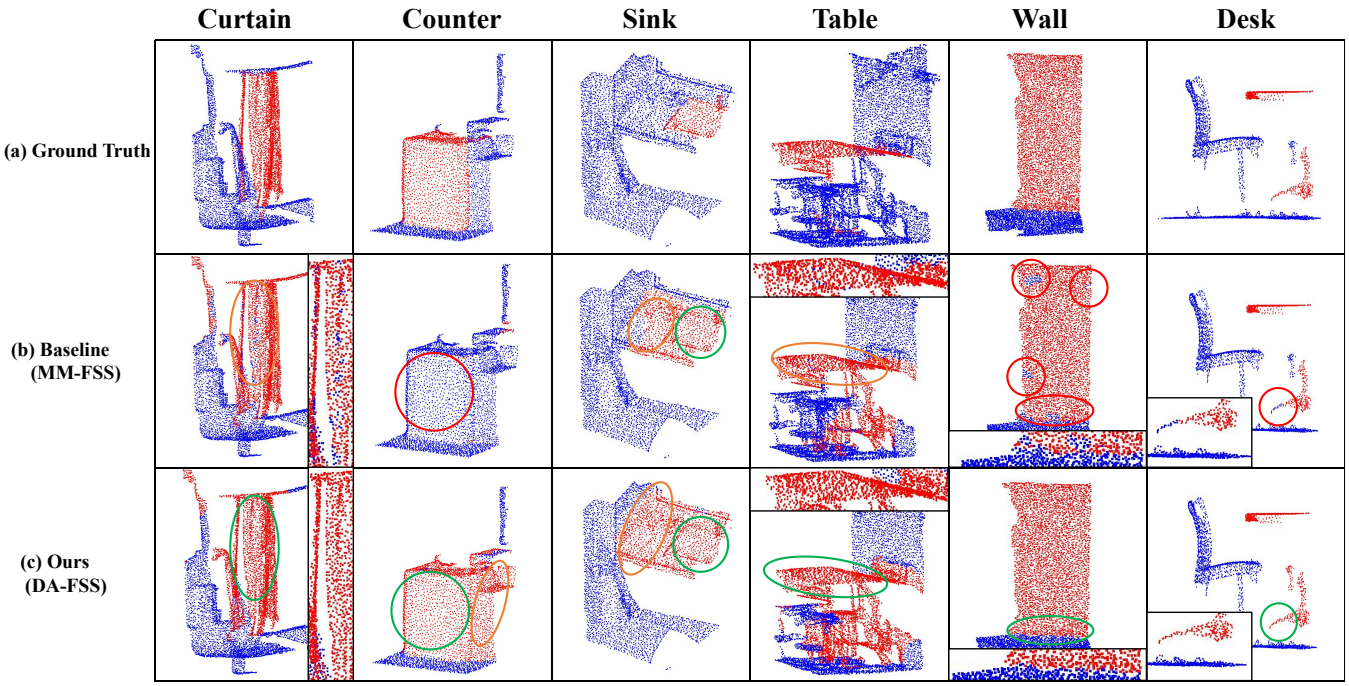

Figure 3: Qualitative comparison between MM-FSS and our proposed DA-FSS in the 1-way 1-shot setting on the ScanNet dataset split 1. The target classes in the columns are curtain, counter, sink, table, wall and desk, respectively. Comparison between (b) Baseline and (c) Ours demonstrates a fundamental trade-off. It is evident that Ours (DA-FSS) exhibits a significant advantage in the geometric completeness of classes and can effectively repair instances of missing points. Moreover, on counter classes, Baseline (MM-FSS) chooses to directly ignore them, whereas Ours (DA-FSS) is capable of recognizing them. However, Ours occasionally exhibits boundary over-segmentation to ensure geometric completeness.

meni *et al.*, 2016] and ScanNet [Dai *et al.*, 2017]. **S3DIS** contains 3D scans from 6 large-scale indoor areas. **ScanNet** consists of 1513 scanned indoor scenes. We adopt the official benchmark splits [Dai *et al.*, 2017], which contain 20 semantic classes. Their semantic classes are split into 2 folds (S^0 and S^1) for cross-validation. Following prior works [An *et al.*, 2025b; Zhao *et al.*, 2021], we maintain a maximum of 20,480 points per block using a 0.02m grid size.

Evaluation Protocol. We follow the standard N-way K-shot episodic evaluation protocol. The primary evaluation metric is the mean Intersection-over-Union (mIoU) over all foreground classes. As in [An *et al.*, 2025b; An *et al.*, 2024], we report results on 1-way 1-shot, 1-way 5-shot, 2-way 1-shot, and 2-way 5-shot settings where the evaluation sets consist of 1,000 episodes per class in the 1-way setting and 100 episodes per class combination in the 2-way setting.

Implementation Details. We adopt the same backbone (Stratified Transformer) [Lai *et al.*, 2022] and feature head architecture as MM-FSS [An *et al.*, 2025b] for fair comparison. Following the standard protocol [An *et al.*, 2025b], we initialize the 3D backbone with pre-trained weights that use the AdamW optimizer, setting a weight decay of 0.01 and a learning rate of 0.006 during pretraining in 100 epochs. Isolating the source of improvement, we choose to download officially released pre-trained weights directly. During the meta-learning phase, the model is trained using the AdamW optimizer with an initial learning rate of 0.0001 and a weight decay of 0.01. For our DA-FSS architecture, we set the number of layers in the Stacked Arbitration Module (SAM) to $N = 1$ for S3DIS and $N = 2$ for ScanNet. For the Decoupled Alignment Module (DAM), the loss weights are set to $\lambda_{PLR} = 0.001$ and $\lambda_{DCR} = 0.5$ based on empirical tuning.

4.2 Comparison with State-of-the-Art

We compare our DA-FSS with previous state-of-the-art methods on S3DIS [Armeni *et al.*, 2016] and ScanNet [Dai *et al.*, 2017], as detailed in Table 1 and Table 2. To ensure a fair assessment, we reproduced the leading baseline MM-FSS [An *et al.*, 2025b] (denoted as MM-FSS†) using its officially released pre-trained weights under our strictly controlled evaluation protocol. In contrast, DA-FSS consistently outperforms the former state-of-the-art across all settings, demonstrating the improved stable generalization brought by decoupling, which enhances the utilization rate of information from pre-trained weights.

Specifically, on the S3DIS dataset, DA-FSS records mIoU increases of **+1.21%** in the 1-way-1-shot setting—which relies heavily on intrinsic geometric generalization. Furthermore, in 1-way-1-shot setting on the ScanNet dataset, DA-

Config.	Architecture Detail	DAM	mIoU (%)
(A) Base	MM-FSS (Fused) [An *et al.*, 2025b]	✗	44.46
(B) Ours	Core Arch. (Decoupled)	✗	44.95
(C) **Ours**	**Full Model (Decoupled+DAM)**	✓	**45.46**

(a) Architectural Effectiveness

Method	FLOPs (G)	Params (M)	mIoU (%)
MM-FSS†	19.06	10.45	44.46
Ours (DA-FSS)	**18.76**	**10.18**	**45.46**
Diff.	**-0.30**	**-0.27**	**+1.00**

(b) Complexity Analysis

Method	mIoU	mAcc(%)	Mechanism Explanation
MM-FSS†	44.46	68.33*	*Catastrophic Miss (High FN)*
Ours	**45.46**	**79.29***	**Object Recovered (High Recall)**
Gain	**+1.0**	**+10.9**	***Solves Semantic Blindness***

(c) **The Decoupling Effect**

Table 3: Comprehensive analysis on ScanNet (1-way 1-shot). We analyze (a) architectural effectiveness, (b) computational complexity, and (c) the specific impact of decoupling. **Note:** DA-FSS achieves a massive surge in mAcc.

FSS has also achieved **+1.00%** improvement over MM-FSS. Notably, DA-FSS surpasses the baseline by **+1.33%** in the 2-way 5-shot setting (see Supp. Material for full tables) that our Decoupled Arbitration mechanism effectively disentangles inter-class semantics under high interference, while simultaneously enforcing intra-class prototype stability through alignment constraints.

Remarkably, beyond standard mIoU metrics, our DA-FSS demonstrates a substantial improvement in **Mean Accuracy (mAcc)** on ScanNet. In settings where mIoU gains appear moderate (e.g., ScanNet 1-way 1-shot), our mAcc often surges by approximately **+10%** compared to the baseline shown in Table 3(c).

Although in some cases ours (DA-FSS) may induce over-segmentation due to its tendency to preserve geometric completeness with high-variance class confusion (which penalizes IoU heavily), this very tendency actually aids the model in more accurately segmenting points that should have been correctly identified. We prioritize semantic completeness over boundary precision, arguing that successfully identifying the target outweighs the cost of minor over-segmentation in 1-way 1-shot setting.

4.3 Ablation Studies

In this section, we conduct a series of ablation studies. Unless specified otherwise, all ablation experiments are performed on the **ScanNet** dataset under the **1-way 1-shot** setting, as this setting is most sensitive to feature quality and generalization capability.

Effectiveness of DA-FSS Architecture. Table 3(a) validates our core hypothesis: "Decoupling" is structurally superior to "Fusion." Comparing (A) and (B), we observe that mIoU increases by approximately **+0.49%** solely through the decoupled design, even when all other components are disabled. This confirms that simply removing the early fusion bottleneck brings gains by mitigating the plasticity-stability conflict. Comparing (B) and (C), we achieved a **+0.51%** improvement after enabling all modules. This proves that while decoupling is essential, the experts must be coordinated via regularization (our DAM) to prevent divergence and achieve optimal alignment thereby optimizing the utilization of information interaction across various modalities..

Complexity Analysis. Finally, Table 3(b) highlights the efficiency of our design. Our DA-FSS achieves a significant performance gain (**+1.00%** mIoU) over the MM-FSS baseline while actually reducing computational overhead (**-0.30** GFLOPs) and parameters (**-0.27** M). This confirms that our gains stem from the superior decoupled architecture, not from increased model capacity.

5 Conclusion

In this paper, we identify a critical conflict in FS-PCS: the "Plasticity-Stability Dilemma," which has undermined the validity of previous progress in fusion-based paradigms. To rectify this issue, we present DA-FSS designed to decouple geometric adaptation from semantic preservation to maximize its adaptability across modalities. DA-FSS combines Parallel Experts to effectively aggregate features, enriching the comprehension of novel concepts from both geometric and semantic perspectives, which are mutually important for recovering missing parts. Furthermore, to ensure expert coordination despite decoupling, we introduce the DAM, which mutually regularizes gradients via prototype alignment. Finally, we introduce the SAM to fuse dual-path features and dynamically synthesize final decisions using boundary-injected guidance. DA-FSS achieves significant improvements over existing methods across all settings, particularly in boosting point-wise accuracy (mAcc). Overall, our research provides valuable insights into the importance of Decoupled Arbitration in FS-PCS and suggests promising directions for future studies.

References

[An *et al.*, 2024] 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.

[An *et al.*, 2025a] Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Junlin Han, Ender Konukoglu, and Serge Belongie. Generalized few-shot 3d point cloud segmentation with vision-language model. In *CVPR*, 2025.

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

[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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1534–1543, 2016.

[Chen *et al.*, 2023] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7020–7030, June 2023.

[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.

[Ding *et al.*, 2023] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7010–7019, June 2023.

[Guo *et al.*, 2023] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.

[He *et al.*, 2024] Qingdong He, Jinlong Peng, Zhengkai Jiang, Kai Wu, Xiaozhong Ji, Jiangning Zhang, Yabiao Wang, Chengjie Wang, et al. Unim-ov3d: Uni-modality open-vocabulary 3d scene understanding with fine-grained feature representation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[Hong *et al.*, 2022] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2022.

[Hong *et al.*, 2023] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 20482–20494. Curran Associates, Inc., 2023.

[Jacobs *et al.*, 1991] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

[Jiang *et al.*, 2024] Li Jiang, Shaoshuai Shi, and Bernt Schiele. Open-vocabulary 3d semantic segmentation with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21284–21294, June 2024.

[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.

[Li *et al.*, 2025] Shuo Li, Fang Liu, Zehua Hao, Xinyi Wang, Lingling Li, Xu Liu, Puhua Chen, and Wenping Ma. Logits deconfusion with clip for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25411–25421, June 2025.

[Liang *et al.*, 2024] Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 32653–32677. Curran Associates, Inc., 2024.

[Nguyen *et al.*, 2024] Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4018–4028, June 2024.

[Ning *et al.*, 2023] Zhenhua Ning, Zhuotao Tian, et al. Boosting few-shot 3d point cloud segmentation via query-guided enhancement. In *ACM MM*, 2023.

[Peng *et al.*, 2023] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–824, June 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 (CVPR)*, July 2017.

[Qi *et al.*, 2017b] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[Qiu *et al.*, 2024] Xinkuan Qiu, Meina Kan, Yongbin Zhou, Yanchao Bi, and Shiguang Shan. Shape-biased cnns are not always superior in out-of-distribution robustness. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2326–2335, January 2024.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Tang *et al.*, 2024] Yiwen Tang, Ray Zhang, Zoey Guo, Xianzheng Ma, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Point-peft: Parameter-efficient fine-tuning for 3d pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 5171–5179, 2024.

[Thomas *et al.*, 2019] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019.

[Tu *et al.*, 2023] Weijie Tu, Weijian Deng, and Tom Gedeon. A closer look at the robustness of contrastive language-image pre-training (clip). In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 13678–13691. Curran Associates, Inc., 2023.

[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.

[Wu *et al.*, 2024a] Jinmeng Wu, Chong Cao, Hao Zhang, Basura Fernando, Yanbin Hao, and Hanyu Hong. Pointtfa: Training-free clustering adaption for large 3d point cloud models. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[Wu *et al.*, 2024b] 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 (CVPR)*, pages 4840–4851, June 2024.

[Xiao *et al.*, 2024] Aoran Xiao, Xiaoqin Zhang, Ling Shao, and Shijian Lu. A survey of label-efficient deep learning for 3d point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9139–9160, 2024.

[Xiong *et al.*, 2024] Guoxin Xiong, Yuan Wang, Zhaoyang Li, Wenfei Yang, Tianzhu Zhang, Xu Zhou, Shifeng Zhang, and Yongdong Zhang. Aggregation and purification: dual enhancement network for point cloud few-shot segmentation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24. International Joint Conferences on Artificial Intelligence Organization (8 2024)*, volume 2, 2024.

[Xu *et al.*, 2024] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer, 2024.

[Xue *et al.*, 2023] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1189, June 2023.

[Xue *et al.*, 2024] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27091–27101, June 2024.

[Yang *et al.*, 2024] Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19823–19832, June 2024.

[Yu *et al.*, 2022] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19313–19322, June 2022.

[Zhang *et al.*, 2022] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022.

[Zhang *et al.*, 2023] Canyu Zhang, Zhenyao Wu, Xinyi Wu, Ziyu Zhao, and Song Wang. Few-shot 3d point cloud semantic segmentation via stratified class-specific attention based transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3410–3417, 2023.

[Zhao *et al.*, 2021] 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.

[Zhao *et al.*, 2022] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022.

[Zhou *et al.*, 2024] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *International Conference on Learning Representations (ICLR)*, 2024.

[Zhu *et al.*, 2023] Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2605–2615, 2023.