EPSegFZ: Efficient Point Cloud Semantic Segmentation for Few- and Zero-Shot Scenarios with Language Guidance

Jiahui Wang[1], **Haiyue Zhu**[2*], **Haoren Guo**[1], **Abdullah Al Mamun**[1], **Cheng Xiang**[1], **Tong Heng Lee**[1]

[1]College of Design and Engineering, National University of Singapore
[2]SIMTech, Agency for Science, Technology and Research (A*STAR)
wjiahui@u.nus.edu, zhu_haiyue@a-star.edu.sg

Abstract

Recent approaches for few-shot 3D point cloud semantic segmentation typically require a two-stage learning process, i.e., a pre-training stage followed by a few-shot training stage. While effective, these methods face overreliance on pre-training, which hinders model flexibility and adaptability. Some models tried to avoid pre-training yet failed to capture ample information. In addition, current approaches focus on visual information in the support set and neglect or do not fully exploit other useful data, such as textual annotations. This inadequate utilization of support information impairs the performance of the model and restricts its zero-shot ability. To address these limitations, we present a novel pre-training-free network, named **E**fficient **P**oint Cloud Semantic **Seg**mentation for **F**ew- and **Z**ero-shot scenarios. Our EPSegFZ incorporates three key components. A **Pro**totype-**E**nhanced **R**egisters **A**ttention (**ProERA**) module and a **D**ual **R**elative **P**ositional **E**ncoding (**DRPE**)-based cross-attention mechanism for improved feature extraction and accurate query-prototype correspondence construction without pre-training. A **L**anguage-**G**uided **P**rototype **E**mbedding (**LGPE**) module that effectively leverages textual information from the support set to improve few-shot performance and enable zero-shot inference. Extensive experiments show that our method outperforms the state-of-the-art method by 5.68% and 3.82% on the S3DIS and ScanNet benchmarks, respectively.

Introduction

Recently, **F**ew-**S**hot **Sem**antic **Seg**mentation (**FS-SemSeg**) for 3D point clouds has gained increasing research interest (Zhao, Chua, and Lee 2021; Zhu et al. 2023; Lai et al. 2022; He et al. 2023a; Zhang et al. 2023a), driven by its potential to efficiently learn from limited data and adapt to unseen categories. However, existing FS-SemSeg approaches heavily rely on fully-supervised pre-trained backbones (Zhao, Chua, and Lee 2021; An et al. 2024b), which can introduce biases due to domain differences between datasets. This is particularly problematic as 3D FS-SemSeg datasets are typically small in size and prone to overfitting. Additionally, the pre-training process is resource-intensive and time-consuming, which limits the practical adoption. Seg-PN (Zhu et al. 2024) attempts to address this challenge by designing a non-parametric encoder. However, it discards high-frequency information to ensure robustness (as shown in Figure 1). We argue that despite potential noise, high-frequency features in point clouds carry essential information for precise object segmentation, particularly edge details. Therefore, developing a method that can both capture high-frequency information effectively and maintain robustness remains a core challenge in FS-SemSeg.

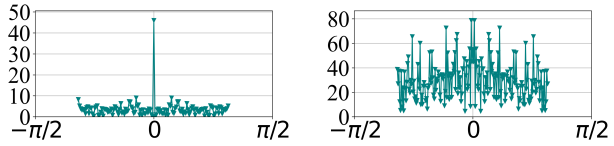


Figure 1: Visualized spectrum (x-axis is frequency, y-axis is amplitute) of embedded features from Seg-PN (**left**) and Ours (**right**) (both are pre-training-free methods). Our latent features are rich and uniform across frequency bands, while Seg-PN overlooks high-frequency components.

To address this challenge, we developed a trainable attention module called **Pro**totype-**E**nhanced **R**egisters **A**ttention (**ProERA**). It utilizes trainable layers to progressively focus on high-frequency information by subtracting low-frequency components and the training process (Xu, Zhang, and Xiao 2019). In addition to incorporating register and prototype tokens, ProERA tackles the foreground-background imbalance, placing greater emphasis on high-frequency details typically found in the minority foreground features. Figure 1 illustrates the frequency spectrum of prototype features processed by our method compared to Seg-PN (Zhu et al. 2024). Our prototype features exhibit richness and uniformity across various frequency bands, whereas Seg-PN features predominantly capture low-frequency information.

While high-frequency information is crucial, achieving a balance between high and low frequencies is essential for a more comprehensive representation. Therefore, low-frequency information should not be overlooked. It is well known that textual embeddings from large pre-trained encoders provide ample low-frequency information (Radford et al. 2021; Bai et al. 2023; Achiam et al. 2023), which can serve as an enhancement in this regard. However, existing

*Corresponding author

models (Zhao, Chua, and Lee 2021; Zhu et al. 2023; Lai et al. 2022; Zhang et al. 2023a; Zhu et al. 2024; An et al. 2024b; Ning et al. 2023; Wang et al. 2025) primarily rely on point labels and disregard textual information. Given the scarcity of labeled data in FS-SemSeg, leveraging textual cues not only supplements low-frequency information to enhance performance but also maximizes the utility of available data.

To this end, we propose **L**anguage-**G**uided **P**rototype **E**mbedding (**LGPE**), which integrates support text annotations to update prototypes within a unified text-3D joint space. This approach enriches low-frequency representations even during the early training stages and helps the model's learning process. Additionally, LGPE enables prototype construction using only text embeddings, introducing a zero-shot capability.

Nonetheless, even with features containing rich high and low frequency information, optimal performance cannot be guaranteed. This is because FS-SemSeg prediction relies on query-prototype correspondence, necessitating additional mechanisms to effectively capture these relationships. Previous approaches have attempted to extract such correlations in various ways: COSeg (An et al. 2024b) designed a cascade network that requires numerous training parameters, while Seg-PN (Zhu et al. 2024) calculated Gram matrices for correlation analysis. However, these methods introduce significant computational overhead through additional parameters and substantial memory consumption.

In response, we propose **D**ual **R**elative **P**ositional **E**ncoding (**DRPE**), which introduces ***no*** extra training parameters. It is the first method to utilize query-prototype relationships within the latent space as Relative Positional Encoding (RPE). Our DRPE process computes the spatial relationships between query and prototype features, incorporating this information as additional input for subsequent cross-attention operations. This approach efficiently captures query-prototype correlations as prior knowledge without resorting to computationally intensive methods. We term this approach DRPE-based cross-attention. The detailed process is illustrated in the Methodology section.

In summary, the main contributions of this work are:

- We propose EPSegFZ, a framework that does not require any pre-training and demonstrates the SOTA performance in 3D FS-SemSeg.
- We developed the ProERA module to enhance extracted features with high-frequency, low-noise information, while its sampling strategies and registers mitigate foreground-background imbalance.
- An LGPE module is proposed to dynamically update prototypes using support textual data, reducing reliance on perfect support point clouds and enabling zero-shot inference.
- A DRPE-based cross-attention is designed as the **first** to use query-prototype spatial relationships as positional signals, accurately establishing correspondence without additional training burden.

Related Work

Point Cloud Semantic Segmentation

Recent research can be generally divided into three main categories. The voxel-based, superpoint-based, and point-based approaches. As documented in (Meng et al. 2019; Choy, Gwak, and Savarese 2019; Wen et al. 2024; Kolodiazhnyi et al. 2024; Zhang, Fei, and Duan 2024), partition the 3D space into regular voxels or superpoints before employing sparse convolutions on them. While these methods demonstrate reasonable performance, they are hindered by imprecise position data resulting from the partition process. The point-based methods (Qi et al. 2017; Thomas et al. 2019; Wu et al. 2024; He et al. 2024; Zhang et al. 2024) directly consider the features and position of each point as inputs, thus preserving natural information. KPConv (Thomas et al. 2019) utilized carefully designed convolution kernels to capture multi-level information. PointNet++ (Qi et al. 2017) trains its Multi-Layer Perceptron (MLP) with point clouds sampled by different strategies to capture global and local information. PointTransformer (Wu et al. 2024) utilizes the attention mechanism to learn point-wise information effectively.

Few-Shot Point Cloud Semantic Segmentation

AttMPTI (Zhao, Chua, and Lee 2021), a groundbreaking approach, utilizes label propagation to uncover connections between prototypes and query points with features extracted by a pre-trained backbone. 2CBR (Zhu et al. 2023) harnesses shared features of support and query to compute bias factors and correct disparities between them. PEFC (Zhang et al. 2023b) implements two specialized components to expand the prototype set and adjust them using query characteristics in a two-way, context-aware fashion. PAPFZS3D (He et al. 2023a) presents a prototype adaptation framework that simultaneously enhances both prototypes and query features. SCAT (Zhang et al. 2023a) introduces a layered, class-specific attention-based transformer, establishing detailed associations between support and query features. COSeg (An et al. 2024b) introduced a correlation memory-based network to reveal the inherent relationship between queries and prototypes. Seg-PN (Zhu et al. 2024) proposes a computationally efficient model employing a non-parametric encoder and correlation-driven support-query interactions to determine point-wise classifications. MM-FSS (An et al. 2024a) uses a multi-modality framework to fuse 2D and 3D information for more details. However, the significant usage of computational resources limited its application in lightweight scenarios.

Methodology

Preliminary

The FS-SemSeg task to be addressed in this work is an N-way K-shot problem (Zhao, Chua, and Lee 2021; Zhu et al. 2024). Consider a support set $\mathcal{S} = \{(\mathbf{P}_s^{n,k}, \mathbf{Y}_s^{n,k})_{k=1}^{K}\}_{n=1}^{N}$, where N is the number of classes, each class consists of K support point cloud samples $\mathbf{P}_s$ with the corresponding labels $\mathbf{Y}_s$. Given the query set $\mathcal{Q} = (\mathbf{P}_q, \mathbf{Y}_q)$, $\mathbf{Y}_q$ is the ground truth which is not available for inference, the goal of this work is to obtain a desirable network $\mathcal{N}_\theta(\cdot)$ and its

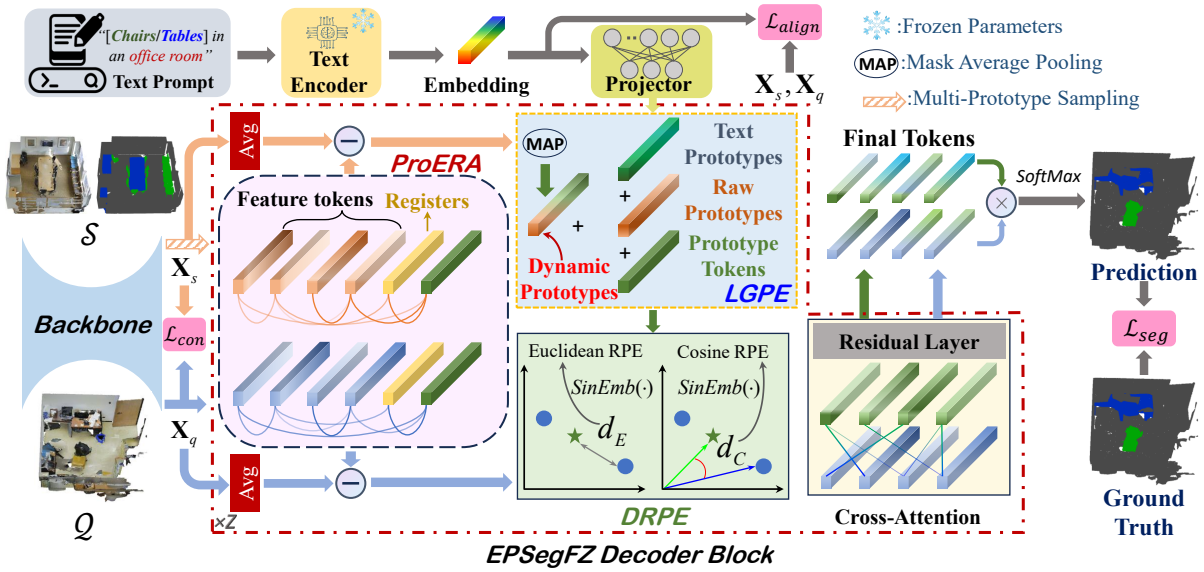
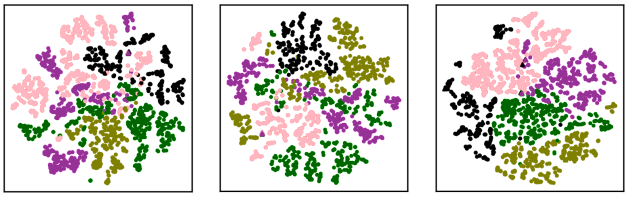


Figure 2: The visualized architecture of our EPSegFZ. A ProERA module first captures high-frequency information and refines the extracted feature. Then, an LGPE module dynamically updates the class prototypes with textual embeddings. After that, a DRPE-based cross-attention properly builds correspondence between prototypes and query features. Finally, the prediction result is obtained by dot production. The red block Avg. represents the average pooling operation.



(a) PAPFZS3D (b) Seg-PN (c) Ours

Figure 3: Visualized t-SNE embedding of feature tokens for prediction. With our LGPE and DRPE, same-class features form a more compact distribution, enhancing the discriminative ability. Colored points represent semantic classes.

parameter θ with the objective:

$$\arg\max_{\theta} \prod_{\mathcal{Q}} p\left(\mathbf{Y}_q \mid \mathcal{N}_\theta\left(\mathbf{P}_q \mid \mathcal{S}\right)\right). \tag{1}$$

Overview

The overall training architecture of our proposed method is illustrated in Figure 2. For each episode in the training, point clouds from the support and query sets are mapped into a latent space using a DGCNN (train from scratch) (Wang et al. 2019). The Multi-Prototype Sampling (MPS) technique (Zhao, Chua, and Lee 2021) is then applied to the support features to derive prototypes. To refine these features and capture high-frequency details, the ProERA module appends learnable registers and prototype tokens to mitigate noise and facilitate query-prototype interaction. Notice that, in each block of the ProERA module, we subtract the average of the input features from the attention output, yielding high-frequency-dominant features. Then, LGPE updates the prototypes with the textual embedding from a language model and adds this to the dynamic prototype obtained by mask average pooling. As shown in Figure 3, t-SNE visualizations compare baselines with our EPSegFZ, demonstrating that our approach achieves a more refined representation space. Next, the DRPE module calculates the correlation between the refined query features and updated prototypes, encoding this information into the subsequent cross-attention module (Qin et al. 2022) for precise correspondence establishment. Finally, we formulate the prediction by computing the similarity based on the dot product between prototypes and query points (Strudel et al. 2021; He et al. 2023a).

Prototype-Enhenced Registers Attention

The attention mechanism has been proven effective in extracting rich semantic and contextual features (Zhao, Chua, and Lee 2021; Dosovitskiy et al. 2021). However, applying attention to all support points can overwhelm the model with irrelevant background data, thus impairing the prototype quality. Besides, early in training, networks prioritize low-frequency information from backgrounds (Xu, Zhang, and Xiao 2019), while our target objects (foreground), represented by fewer points, contain higher-frequency data due to their complex geometry and semantics that might be overlooked.

Considering this, we replace support points with multi-prototypes sampled from support data and add registers for both query and prototype tokens. This approach reduces computational load and mitigates the potential influence of a large number of irrelevant background points. Moreover, self-attention usually acts as a low-pass filter (Wang and

et.al. 2022), emphasizing low-frequency over high-frequency. Therefore, we use the average pooling operation to pre-calculate the overall low-frequency information and subtract such a feature from the output to produce features containing sufficient high-frequency information. Figure 4 illustrates the similarity map between query features and registers. It suggests that the registers prioritize distinct areas within the scene: one register directs attention toward the background and object-free area, whereas the other emphasizes the area that notably contains multiple objectives. This method also mitigates the background-foreground imbalance, which implicitly addresses low-frequency vs. high-frequency imbalance during training, thereby enhancing the representational capability.

Specifically, given a support point cloud $\mathbf{P}_s$ and a query point cloud $\mathbf{P}_q$, each consisting of M points, their features are first obtained with a backbone and denoted as $\mathbf{X}_t \in \mathbb{R}^{M\times D}; t \in \{s, q\}$ where D is the embedded dimension. In our ProERA module, we append extra n_r learnable tokens, denoted as $\mathbf{r}_t \in \mathbb{R}^{n_r\times D}$, as shown in the Figure 2. Besides, to strengthen the interactions between prototypes and these features, we additionally append prototype tokens behind registers. The multi-prototype $\mathbf{X}_p$ and raw prototype token $\mathbf{p}_{raw}$ of the c-th category are denoted as:

$$\mathbf{X}_p = \text{MPS}\left(\mathbf{X}_s\right); \mathbf{p}^c_{raw} = \frac{1}{n_p}\sum_{n_p}(\mathbf{X}_p \times \mathbb{1}(\mathbf{Y}_p = c)), \quad (2)$$

where n_p is a hyperparameter determining the number of sampled prototypes for each class, MPS$(\cdot)$ is the multi-prototype sampling algorithm (Zhao, Chua, and Lee 2021), and $\mathbf{Y}_p$ is the label of the prototype. The function $\mathbb{1}(\cdot)$ is the binary label indicator that outputs 1 when its variable is true. Consider the i-th decoder block ($i \in [1, Z]$), with the residual layer Res$(\cdot)$ and self-attention layer SA$(\cdot)$ the output of our ProERA module is:

$$\tilde{\mathbf{X}}^i_j = \text{Res}\left(\text{SA}([\hat{\mathbf{X}}^{i-1}_j; \mathbf{r}_j; \hat{\mathbf{p}}^{i-1}])\right) - \frac{1}{n_j}\sum_{n_j}\hat{\mathbf{X}}^{i-1}_j, \quad (3)$$

where $[;;]$ is the concatenation operation, $j \in \{p, q\}$ denotes equal operation to query and support. $\hat{\mathbf{p}}^{i-1}$ and $\hat{\mathbf{X}}^{i-1}_j$ denote the prototype tokens and feature tokens from the last block, respectively; their definition will be illustrated in the following section. For the first decoder block, the relative inputs are $\hat{\mathbf{X}}^0_j = \mathbf{X}_j$ and $\hat{\mathbf{p}}^0 = \mathbf{p}_{raw}$.

Language-Guided Prototype Embedding

Due to the limited data available in FS-SemSeg, the declined representation of prototypes caused by imperfect support point clouds hinders the learning process. Moreover, if the backbone is not pre-trained, the prototypes in the early stage of training are not discriminative. These findings prompt us to reduce the model's reliance on solely visual support information. We propose an LGPE module to optimize the prototypes and enable the network to conduct zero-shot inference using support text labels.

Our LGPE module receives the embedding feature from the text encoder of a pre-trained CLIP (Radford et al. 2021) that is prompted with corresponding support classes to form text prototypes:

$$\mathbf{p}^c_{text} = \text{Proj}\left(\mathbf{T}^c\right), \quad (4)$$

where $\mathbf{T}^c$ refers to a text embedding from the CLIP text encoder for the c-th class, and Proj$(\cdot)$ is a projection network. Take the i-th block in the decoder as an example, the dynamic prototypes are obtained with the current multi-prototype $\tilde{\mathbf{X}}^i_p$ and mask average pooling, i.e., $\mathbf{p}^{i,c}_{dyn} = \frac{1}{n_p}\sum_{n_p}\tilde{\mathbf{X}}^{i,c}_p$.

Subsequently, the updated prototypes represent a blend of previous prototype tokens, the raw prototypes, the text prototypes, and the dynamic prototypes, as indicated by:

$$\mathbf{p}^i = \lambda_1\tilde{\mathbf{p}}^i + \lambda_2\mathbf{p}_{raw} + \lambda_3\mathbf{p}^i_{dyn} + \lambda_4\mathbf{p}_{text}, \quad (5)$$

where $\tilde{\mathbf{p}}^i$ is obtained from $\tilde{\mathbf{X}}^i_p$ with mask average pooling. Considering that text embedding is already well-established while visual embedding starts with a weaker representation, their fusion into the prototype should dynamically adjust over time. Initially, the model relies more on the text embedding to provide strong guidance, as the visual embedding has not yet learned meaningful features. As training progresses and the visual representation improves, its weight gradually increases, allowing the model to shift towards a more balanced fusion of both modalities. This mechanism ensures a smooth transition from text-driven supervision to a learned visual-text alignment. Formally, we define the weight of the text embedding as $\lambda_4(t) = \lambda^*_4 e^{-0.5t}$, where λ^*_4 is a predefined final weight. Meanwhile, the weights for visual embeddings, $\lambda_i(t), t \in [1, 3]$, follow an increasing function $\lambda_i(t) = \lambda^*_i(1 - e^{-0.5t})$, ensuring a gradual adaptation that shifts the prototype's reliance from text to learned visual features over time. This processing not only provides favorable prototypes but also mitigates the influence of the random-initialized backbone in the early stage of training. Figure 4 illustrates that employing the updated prototypes contributes to establishing preferable correspondence, thereby facilitating more precise predictions.

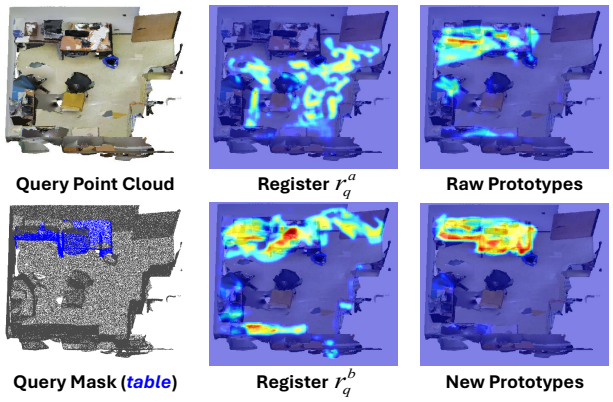


Figure 4: Visualized heatmaps of query-registers and query-prototypes similarities. The distinct focused region of registers helps the model differentiate between object-related and object-less areas. The updated prototypes effectively correlate with the query object, whereas the raw prototypes lack sufficient focus on the target object.

Dual Relative Positional Encoding

Our DRPE module is the first to incorporate the inter-cloud correlation (query-prototype) in the latent space as a positional encoding signal in cross-attention for FS-SemSeg tasks. Specifically, the Euclidean distance between the j-th point in the query point cloud and the c-th prototype in the i-th block is $d_E^{i,j,c}$. Hence, the distance between all query points and the c-th prototype is a vector that can be expressed as $\mathbf{d}_E^{i,c} \in \mathbb{R}^{M\times 1}$. This vector is then used as indices to calculate the encoding value $\mathbf{R}_E^{i,c} \in \mathbb{R}^{M\times(N+1)\times D}$ with the sinusoidal positional encoding function (Vaswani et al. 2017a). Simultaneously, the cosine value of the angle formed between a query vector and a prototype vector is calculated and indicated as $d_C^{i,j,c}$. This distance is also encoded with the same sinusoidal function and denoted as $\mathbf{R}_C^i \in \mathbb{R}^{M\times(N+1)\times D}$. The DRPE value is an element-wise addition of the two results, $\mathbf{R}^i = \mathbf{R}_C^i + \mathbf{R}_E^i$., With the calculated DRPE, we can formulate a DRPE-based decoder; its visualized architecture is shown in the Appendix.

For prediction, we follow (Strudel et al. 2021) to conduct the pipeline. The query feature $\tilde{\mathbf{X}}_q^i$ and the updated prototypes $\mathbf{p}^i$ are fed into a cross-attention module, then processed sequentially. The final query feature token $\hat{\mathbf{X}}_q^Z$ and prototypes $\hat{\mathbf{p}}^Z$ are the output of the Z-th decoder which are used to generate the prediction result

$$\hat{\mathbf{Y}}_q = \text{SoftMax}(\frac{\hat{\mathbf{X}}_q^Z}{||\hat{\mathbf{X}}_q^Z||_2} \cdot \frac{\hat{\mathbf{p}}^Z}{||\hat{\mathbf{p}}^Z||_2}). \tag{6}$$

Each row in $\hat{\mathbf{Y}}_q$ represents the probability that the corresponding point belongs to each support class. Detailed calculation process within the cross-attention can be accessed in the appendix.

Loss Functions

To achieve better performance, the loss function must be carefully designed. Since our backbone is not pre-trained, we propose a foreground-consistency loss to enhance its feature representation, as shown in Equation 7. It encourages positive pairs, features from the same foreground region, to be closer in the embedding space while pushing negative pairs apart. By focusing on foreground consistency, our method helps the backbone learn more discriminative features while mitigating instability caused by the lack of pre-training. Thus, it is defined as:

$$\mathcal{L}_{con} = \text{InfoNCE}(\mathbf{x}_q, \mathbf{x}_s) \tag{7}$$

where InfoNCE$(\cdot)$ is an InfoNCE-based (He et al. 2020) constrastive loss, τ is a temperature hyperparameter. The detailed formulation of $\mathcal{L}_{con}$ is illustrated in the appendix.

Building on this foundation, we introduce a foreground-aware alignment loss that serves two key purposes: reducing the gap between visual and text embeddings to enhance the text-visual joint space while also providing crucial guidance during early training stages. This loss function minimizes the cross-entropy between text-visual similarity and text labels, encouraging better alignment across modalities. It ensures that the features of the foreground objects closely match their textual descriptions, leading to a more meaningful relationship between image and text and improved performance in multimodal tasks. Mathematically, the loss can be written as:

$$\mathcal{L}_{align} = \frac{1}{N}\sum_{c=1}^{N} \text{CE}\left(\text{SoftMax}(W\mathbf{p}_{raw}^c \cdot \mathbf{T}^c), c\right), \tag{8}$$

where W is a learnable matrix, CE$(\cdot)$ is the cross-entropy loss. $\mathbf{p}_{raw}^c$ is the visual prototype for the c-th foreground, and $\mathbf{p}_{text}$ is a feature map that represents text embeddings of all foreground objects.

Lastly, the main segmentation loss supervises all modules by minimizing the cross-entropy between predicted $\hat{\mathbf{Y}}_q$ and the ground-truth $\mathbf{Y}_q$: $\mathcal{L}_{seg} = \text{CE}(\mathbf{Y}_q, \hat{\mathbf{Y}}_q)$. Therefore, our final loss in training is: $\mathcal{L} = \mathcal{L}_{seg} + \lambda_{con}\mathcal{L}_{con} + \lambda_{align}\mathcal{L}_{align}$

Model	#Params.	GFLOPs	Time(s)	Δ (m-IoU)
AttMPTI	372.19K	2.60	0.46	-19.31
PAPFZS3D	2.46M	3.07	0.62	-13.63
Seg-PN	241.67K	1.95	0.32	-8.24
COSeg	7.69M	9.71	1.35	+0.05
Ours	2.02M	2.11	0.36	-

Table 1: Efficiency analysis on S3DIS based on model complexity, FLOPs, and inference time.

Experiments

Implementation Details

Data preparation. We construct our FS-SemSeg tasks on the S3DIS (Armeni et al. 2017) and the ScanNet (Dai et al. 2017) datasets. As in previous works (Zhao, Chua, and Lee 2021; He et al. 2023b; Wang et al. 2023), we adopt the preprocessing methodology proposed in (Qi et al. 2017), which involves splitting rooms into smaller units. For training, we randomly selected 2048 points per unit. We partition each dataset into two distinct subsets, S^0 and S^1, as prior methodologies (Zhao, Chua, and Lee 2021; He et al. 2023b), where one subset exclusively serves for inference, and the other is designated for training.

Training and Inference. As in previous works (He et al. 2023b; Zhang et al. 2023b), episodic learning is adopted in the training and testing. Our total iteration number is set to 30,000. To obtain a meaningful visual representation in the early training stage, we set a relatively large learning rate for the backbone to enable faster updates. We also assign a smaller decay step, which decreases more rapidly than the learning rate of other modules. Moreover, we implement a random shuffle of the point order in the query and support the data for fairness. During inference, 100 episodes are randomly selected. Text embeddings are obtained in advance and saved locally as a dictionary to mitigate computational burden. For detailed hyperparameter settings, please refer to the appendix.

Method	2-way 1-shot				2-way 5-shot				3-way 1-shot				3-way 5-shot			
	S^0	S^1	mean	Δ	S^0	S^1	mean	Δ	S^0	S^1	mean	Δ	S^0	S^1	mean	Δ
Fine-Tuning	36.34	38.79	37.57	-35.85	56.49	56.99	56.74	-18.27	30.05	32.19	31.12	-34.81	46.88	47.57	47.23	-21.05
ProtoNet	48.39	49.98	49.19	-25.89	57.34	63.22	60.28	-15.73	40.81	45.07	42.94	-22.99	49.05	53.42	51.24	-17.04
AttMPTI	53.77	55.94	54.86	-18.56	61.67	67.02	64.35	-11.66	45.18	49.27	47.23	-18.70	54.92	56.79	55.86	-12.42
CWT	52.14	57.86	55.00	-18.42	61.64	66.48	64.06	-11.95	-	-	-	-	-	-	-	-
2CBR	55.89	61.99	58.94	-14.48	63.55	67.51	65.53	-10.48	46.51	53.91	50.21	-15.72	55.51	58.07	56.79	-11.49
SCAT	54.92	56.74	55.83	-17.59	64.24	69.03	66.63	-9.38	-	-	-	-	-	-	-	-
PEFC	55.09	59.63	57.36	-16.06	65.47	70.84	68.16	-7.85	49.15	54.69	51.92	-14.01	62.56	63.21	62.89	-5.39
QGE	58.85	60.29	59.57	-13.85	66.56	**79.46**	69.01	-7.00	-	-	-	-	-	-	-	-
PAPFZS3D	59.45	66.08	62.76	-10.66	65.40	70.30	67.85	-8.16	48.99	56.57	52.78	-13.15	61.27	60.81	61.04	-7.24
Seg-PN	64.84	67.98	66.41	-7.01	67.63	71.48	69.36	-6.65	60.12	63.22	61.67	-4.26	62.58	64.53	63.56	-4.72
SDSimPoint	68.73	70.61	69.67	-3.75	72.12	72.72	72.42	-3.59	62.28	62.11	62.19	-3.74	65.17	66.10	65.64	-2.64
Ours+Point-NN	72.31	**74.20**	73.26	-0.16	75.26	75.94	75.60	-0.41	**65.63**	66.09	65.86	-0.07	67.94	**68.55**	68.25	-0.03
Ours+DGCNN	**73.08**	73.75	**73.42**	-	**75.90**	76.11	**76.01**	-	65.58	**66.27**	**65.93**	-	**68.30**	68.25	**68.28**	-

Table 2: Evaluation result on the S3DIS dataset using mean-IoU criteria (%). The best result of each column is highlighted with **bold font**, and the second best is noted with underline.

Method	2-way 1-shot				2-way 5-shot				3-way 1-shot				3-way 5-shot			
	S^0	S^1	mean	Δ	S^0	S^1	mean	Δ	S^0	S^1	mean	Δ	S^0	S^1	mean	Δ
Fine-Tuning	31.55	28.94	30.25	-38.59	42.71	37.24	39.98	-30.66	23.99	19.10	21.55	-45.48	34.93	28.10	31.52	-38.21
ProtoNet	33.92	30.95	32.44	-36.40	45.34	42.01	43.68	-26.96	28.47	26.13	27.30	-39.73	37.36	34.98	36.17	-33.56
AttMPTI	42.55	40.83	41.69	-27.15	54.00	50.32	52.16	-18.48	35.23	30.72	32.98	-34.05	46.74	40.80	43.77	-25.96
CWT	42.33	41.78	42.05	-26.79	55.60	53.77	56.48	-14.16	-	-	-	-	-	-	-	-
2CBR	50.73	47.66	49.20	-19.64	52.35	47.14	49.75	-20.89	47.00	46.36	46.68	-20.35	45.06	39.47	42.27	-27.46
SCAT	45.24	45.90	45.57	-23.27	55.38	57.11	56.24	-14.40	-	-	-	-	-	-	-	-
PEFC	45.31	44.86	45.09	-23.75	56.26	54.06	55.16	-15.48	38.78	36.13	37.46	-29.57	51.72	46.05	48.89	-20.84
QGE	43.10	46.79	44.95	-23.89	51.91	57.21	54.56	-16.08	-	-	-	-	-	-	-	-
PAPFZS3D	57.08	55.94	56.51	-12.33	64.55	59.64	62.10	-8.54	55.27	55.60	55.44	-11.59	59.02	53.16	56.09	-13.64
Seg-PN	63.15	64.32	63.74	-5.10	67.08	69.05	68.07	-2.57	61.80	65.34	63.57	-3.46	62.94	68.26	65.60	-4.13
SDSimPoint	65.21	65.18	65.19	-3.65	68.20	68.49	68.35	-2.29	63.30	63.86	63.83	-3.20	65.04	66.27	65.66	-4.07
Ours+Point-NN	65.71	66.01	65.86	-2.98	67.94	69.22	68.58	-2.06	64.32	65.57	64.95	-2.08	66.32	68.74	67.53	-2.20
Ours+DGCNN	**69.43**	**68.25**	**68.84**	-	**71.31**	**69.97**	**70.64**	-	**67.19**	**66.86**	**67.03**	-	**69.62**	**69.83**	**69.73**	-

Table 3: Evaluation result on the ScanNet dataset using mean-IoU criteria (%). The best result of each column is highlighted with **bold font**, and the second best is noted with underline.

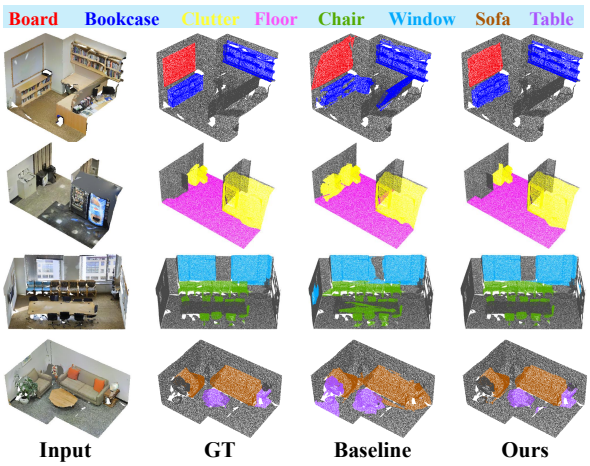


Figure 5: Visualized segmentation result on S3DIS dataset. Our method performs better in segmentation accuracy than the baseline.

Experimental Results

Table 1 shows the efficiency analysis (Δ represents the gap between ours and the corresponding method). Our method demonstrates efficiency with comparable FLOPs and inference time, given the enhanced accuracy, making it a competitive approach. Table 2 and Table 3 present the performance of the different models in the S3DIS (Armeni et al. 2017) and ScanNet (Dai et al. 2017) datasets. We tested our approach with the trainable DGCNN backbone (Wang et al. 2019) and the nonparametric Point-NN backbone (Zhang et al. 2023c) used in Seg-PN (Zhu et al. 2024). Our proposed approach demonstrates a substantial improvement in all FS-SemSeg settings and outperforms previous SOTA by **5.68%** and **3.82%** m-IoU on S3DIS (Armeni et al. 2017) and ScanNet (Dai et al. 2017), respectively. We introduce each compared baselines in the appendix. Figure 5 depicts the visualized segmentation results of our method and baseline (Seg-PN (Zhu et al. 2024)); in contrast to the baseline, our EPSegFZ exhibits clearer segmentation on edges, benefiting from ample high-frequency information.

Zero-Shot Scenario Evaluation

We conduct zero-shot experiments with CLIP (Radford et al. 2021) and word2vec (Mikolov et al. 2013) as language models. PAPFZS3D (He et al. 2023a) uses prototype-based regression with text embeddings applied post-interaction, limiting their influence on predictions.

In contrast, our zero-shot prototypes are initialized from

Embed	Method	2-way 1-shot	2-way 5-shot	3-way 1-shot	3-way 5-shot
word2vec	3DGenZ	34.93	36.12	23.08	27.52
word2vec	PAPFZS3D	59.98	63.54	48.91	55.62
CLIP	PAPFZS3D	61.09	64.91	50.18	59.10
CLIP	**Ours**	**63.84**	**65.43**	**55.62**	**60.04**

Table 4: Zero-Shot evaluation result on the S3DIS dataset using mean-IoU criteria (%). The best result of each column is highlighted with **bold font**.

ID	ProERA	LGPE	DRPE	Result	Δ
I				31.55	-41.53
II			✓	64.84	-8.24
III		✓		60.22	-12.86
IV	✓			59.27	-13.81
V		✓	✓	70.48	-2.60
VI	✓	✓		69.35	-3.73
VII	✓		✓	70.17	-2.91

ID	$\mathbf{p}_{text}$	$\mathbf{p}_{dyn}$	$\mathbf{p}_{raw}$	Result	Δ
VIII				68.71	-4.37
IX	✓			71.49	-1.59
X		✓		70.30	-2.78
XI			✓	69.51	-3.57
XII	✓	✓		71.75	-1.33
XIII	✓		✓	71.08	-2.00
XIV		✓	✓	70.56	-1.52
XV	✓	✓	✓	**73.08**	-

Table 5: Ablation study of model components (**upper**) and prototypes (**lower**) on S3DIS S^0 with 2-way 1-shot.

the language model and refined through direct interaction, enabling better alignment between query tokens and text features. This preserves more semantic information for final similarity matching. We focus on validating zero-shot feasibility and do not compare with large-scale models like SegPoint (He et.al 2024) due to differences in training resources and data scale. Note that this experiment aims to validate the feasibility of zero-shot inference with our method. We do not compare against large-scale models such as SegPoint (He et.al 2024), as there exists a significant disparity in training resources and the scale of utilized data. The results in Table 4 demonstrate the efficacy and superiority of our proposed approach.

Ablation Study

Table 5 highlights the effectiveness of each model component and each type of prototype, demonstrating our approach's ability to enhance prediction accuracy. The results show that dynamic prototypes have the most significant impact on performance. Table 6 confirms that both $\mathcal{L}_{con}$ and $\mathcal{L}_{align}$ significantly improve performance. To evaluate the effectiveness of our proposed DRPE, we conduct an ablation study comparing it with learnable positional encoding (*Learn PE* in the table) (Wang et al. 2022) and sinusoidal (Vaswani et al. 2017b) positional encodings (*Sin PE* in the table). As shown in Table 6, DRPE demonstrates clear advantages. Given that 3D point coordinates already provide accurate spatial positions, DRPE further enriches the representation by fusion query-support information.

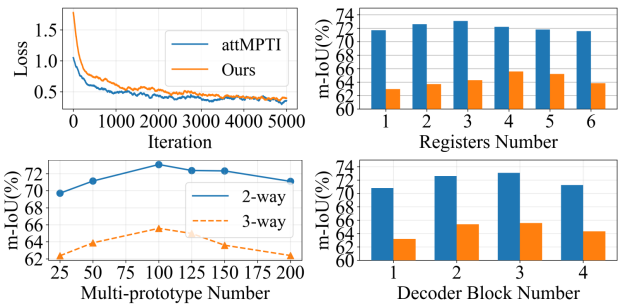


Figure 6: Visualized training loss curve (**upper left**). Ablation study results on number of registers (**upper right**), decoder block (**lower right**), and multi-prototype (**lower left**).

ID	Components	Result	Δ
I	$w/o\ \mathcal{L}_{align}$	69.46	-4.24
II	$w/o\ \mathcal{L}_{con}$	69.30	-4.50
III	$w/o\ \mathcal{L}_{align}$+$\mathcal{L}_{con}$	68.55	-5.25
IV	Sin PE (Vaswani et al. 2017b)	69.94	-3.14
V	Learn PE (Wang et al. 2022)	71.22	-1.86
VI	**Ours**	**73.08**	-

Table 6: Ablation study of losses (upper) and positional encoding (lower) on S3DIS S^0.

Figure 6 presents loss curves of attMPTI and our model over the first 5,000 iterations, showing similar convergence speeds. Our model achieves stable loss values comparable to a fully-supervised pre-trained method. Figure 6 also examines performance variations with different register and decoder block counts in 2-way 1-shot and 3-way 1-shot settings. As observed in (Darcet et al. 2024), register count requires careful tuning. We find $N+1$ registers work best for N-way K-shot tasks, and 100 prototypes with 3 blocks have the best performance-efficiency trade-off.

Conclusion

We propose EPSegFZ, a pre-training-free 3D SemSeg model targeted for few-shot and zero-shot scenarios. A ProERA module is developed to enable the network to capture high-frequency features with less noise. Accurate query-prototype correspondence can be established using the proposed DRPE-based cross-attention. An LGPE module is designed to update prototypes with textual support information, fully exploit available data, reduce reliance on perfect visual information, and empower zero-shot inference. Furthermore, we designed a foreground-consistency alignment loss and a foreground-aware contrastive loss to effectively supervise the language-vision alignment and feature extraction.

Acknowledgements

This research is supported by the National University of Singapore under the NUS College of Design and Engineering Industry-focused Ring-Fenced PhD Scholarship programme. It is also supported by the National Research Foundation (NRF) "Centre for Advanced Robotics Technology Innovation (CARTIN)", and the National Robotics Programme (NRP) 2.0 funding initiative "Domain-specific Robotics Foundation Models for Manufacturing (DS-RFM). The authors would like to acknowledge useful discussions with Dr. Bruce Engelmann from Hexagon, Manufacturing Intelligence Division, Simufact Engineering GmbH.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

An, Z.; Sun, G.; Liu, Y.; Li, R.; Wu, M.; Cheng, M.-M.; Konukoglu, E.; and Belongie, S. 2024a. Multimodality Helps Few-Shot 3D Point Cloud Semantic Segmentation. arXiv:2410.22489.

An, Z.; Sun, G.; Liu, Y.; Liu, F.; Wu, Z.; Wang, D.; Van Gool, L.; and Belongie, S. 2024b. Rethinking Few-shot 3D Point Cloud Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3996–4006.

Armeni, I.; Sax, A.; Zamir, A. R.; and Savarese, S. 2017. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Choy, C.; Gwak, J.; and Savarese, S. 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 3075–3084.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Darcet, T.; Oquab, M.; Mairal, J.; and Bojanowski, P. 2024. Vision Transformers Need Registers. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9729–9738.

He, S.; Ding, H.; Jiang, X.; and Wen, B. 2024. Segpoint: Segment any point cloud via large language model. In *European Conference on Computer Vision*, 349–367. Springer.

He, S.; Jiang, X.; Jiang, W.; and Ding, H. 2023a. Prototype Adaption and Projection for Few- and Zero-shot 3D Point Cloud Semantic Segmentation. *IEEE Transactions on Image Processing*.

He, S.; Jiang, X.; Jiang, W.; and Ding, H. 2023b. Prototype Adaption and Projection for Few- and Zero-Shot 3D Point Cloud Semantic Segmentation. *IEEE Transactions on Image Processing*, 32: 3199–3211.

He et.al, S. 2024. SegPoint: Segment Any Point Cloud via Large Language Model. In *ECCV*.

Kolodiazhnyi, M.; Vorontsova, A.; Konushin, A.; and Rukhovich, D. 2024. Oneformer3d: One transformer for unified point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20943–20953.

Lai, L.; Chen, J.; Zhang, C.; Zhang, Z.; Lin, G.; and Wu, Q. 2022. Tackling background ambiguities in multi-class few-shot point cloud semantic segmentation. *Knowledge-Based Systems*, 253: 109508.

Meng, H.-Y.; Gao, L.; Lai, Y.-K.; and Manocha, D. 2019. Vv-net: Voxel vae net with group convolutions for point cloud segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8500–8508.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. ArXiv: 1301.3781.

Ning, Z.; Tian, Z.; Lu, G.; and Pei, W. 2023. Boosting few-shot 3d point cloud segmentation via query-guided enhancement. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1895–1904.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 652–660.

Qin, Z.; Yu, H.; Wang, C.; Guo, Y.; Peng, Y.; and Xu, K. 2022. Geometric Transformer for Fast and Robust Point Cloud Registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11143–11152.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7262–7272.

Thomas, H.; Qi, C. R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; and Guibas, L. J. 2019. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international Conference on Computer Vision and Pattern Recognition (CVPR)*, 6411–6420.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017a. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017b. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, G.; Lu, Y.; Cui, L.; Lv, T.; Florencio, D.; and Zhang, C. 2022. A simple yet effective learnable positional encoding method for improving document transformer model. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, 453–463.

Wang, J.; Zhu, H.; Guo, H.; Al Mamun, A.; Xiang, C.; de Silva, C. W.; and Lee, T. H. 2025. SDSimPoint: Shallow–Deep Similarity Learning for Few-Shot Point Cloud Semantic Segmentation. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, J.; Zhu, H.; Guo, H.; Mamun, A. A.; Xiang, C.; and Lee, T. H. 2023. Few-Shot Point Cloud Semantic Segmentation via Contrastive Self-Supervision and Multi-Resolution Attention. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2811–2817.

Wang, P.; and et.al. 2022. Anti-Oversmoothing in Deep Vision Transformers via the Fourier Domain Analysis: From Theory to Practice. In *International Conference on Learning Representations (ICLR)*.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (ToG)*, 38(5): 1–12.

Wen, K.; Zhang, N.; Li, G.; and Gao, W. 2024. MPVNN: multi-resolution point-voxel non-parametric network for 3d point cloud processing. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Wu, X.; Jiang, L.; Wang, P.-S.; Liu, Z.; Liu, X.; Qiao, Y.; Ouyang, W.; He, T.; and Zhao, H. 2024. Point Transformer V3: Simpler Faster Stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4840–4851.

Xu, Z.-Q. J.; Zhang, Y.; and Xiao, Y. 2019. Training behavior of deep neural network in frequency domain. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26*, 264–274. Springer.

Zhang, C.; Wu, Z.; Wu, X.; Zhao, Z.; and Wang, S. 2023a. 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, 3410–3417.

Zhang, H.; Wang, C.; Yu, L.; Tian, S.; Ning, X.; and Rodrigues, J. 2024. Pointgt: A method for point-cloud classification and segmentation based on local geometric transformation. *IEEE Transactions on Multimedia*.

Zhang, Q.; Wang, T.; Hao, F.; Wu, F.; and Cheng, J. 2023b. Prototype expansion and feature calibration for few-shot point cloud semantic segmentation. *Neurocomputing*, 558: 126732.

Zhang, R.; Wang, L.; Guo, Z.; Wang, Y.; Gao, P.; Li, H.; and Shi, J. 2023c. Parameter is not all you need: Starting from non-parametric networks for 3d point cloud analysis. *arXiv preprint arXiv:2303.08134*.

Zhang, S.; Fei, X.; and Duan, Y. 2024. GeoAuxNet: Towards Universal 3D Representation Learning for Multi-sensor Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20019–20028.

Zhao, N.; Chua, T.-S.; and Lee, G. H. 2021. Few-shot 3d point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8873–8882.

Zhu, G.; Zhou, Y.; Yao, R.; and Zhu, H. 2023. Cross-class bias rectification for point cloud few-shot segmentation. *IEEE Transactions on Multimedia*, 25: 9175–9188.

Zhu, X.; Zhang, R.; He, B.; Guo, Z.; Liu, J.; Xiao, H.; Fu, C.; Dong, H.; and Gao, P. 2024. No Time to Train: Empowering Non-Parametric Networks for Few-shot 3D Scene Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3838–3847.