

# Few-shot 3D Point Cloud Semantic Segmentation

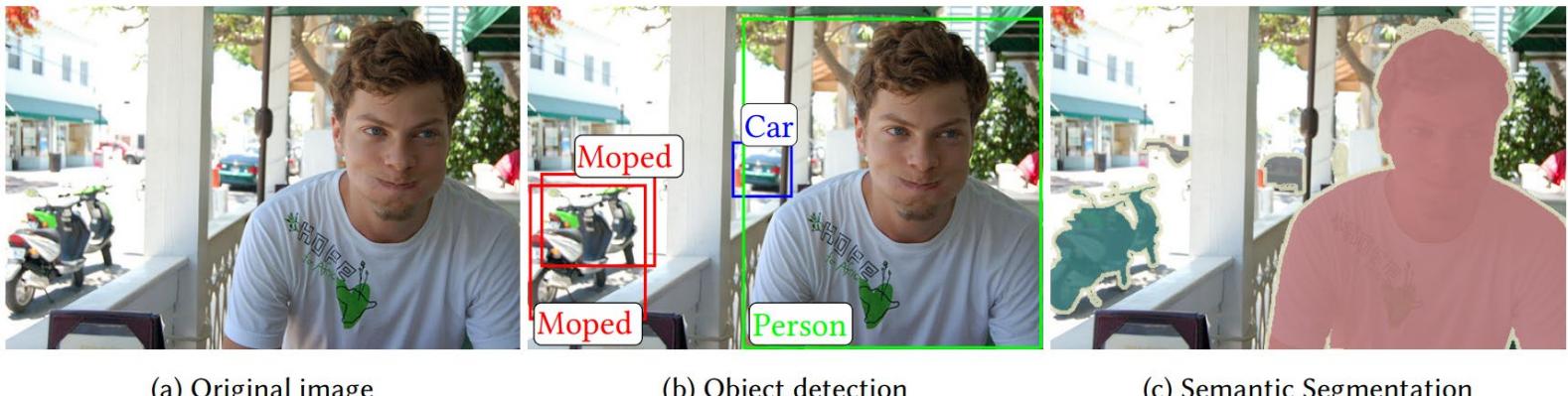
Yun Liu

Professor, Nankai University

- 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”. **IEEE CVPR, 2024**.
- 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”. **ICLR 2025**.
- 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”. **IEEE CVPR, 2025**.

# Image Semantic Segmentation

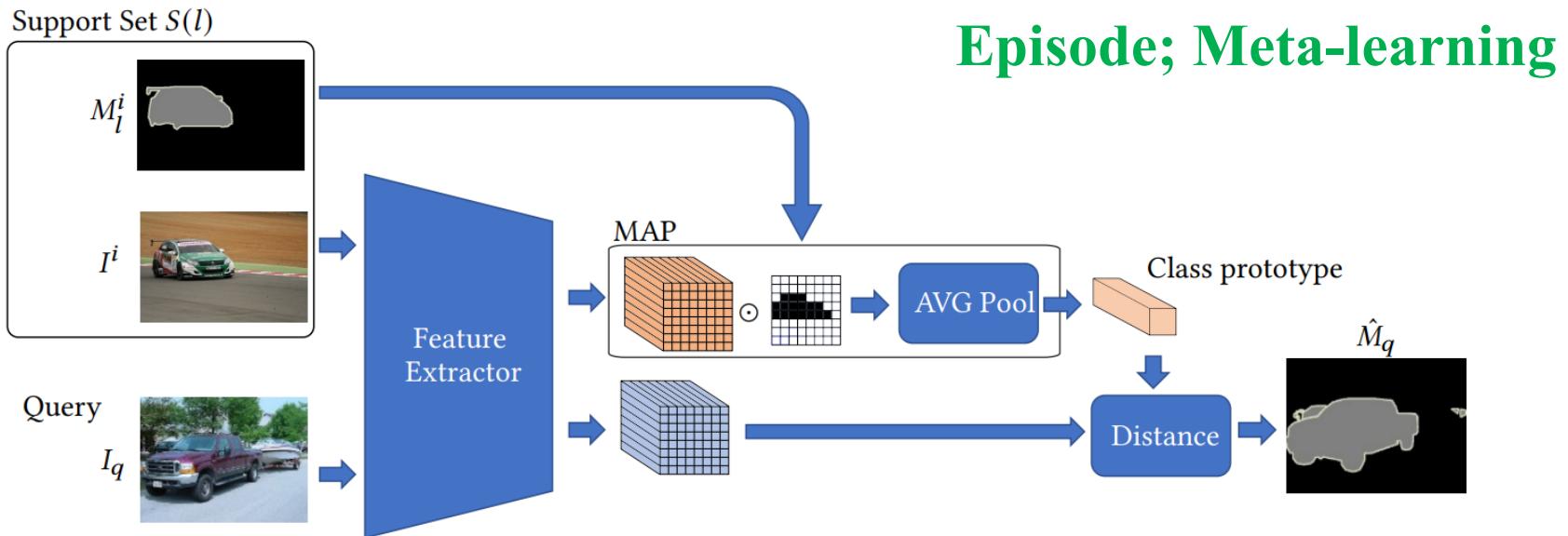
- Training fully-supervised image semantic segmentation models requires large-scale datasets with pixel-wise annotations. However, creating and labeling such datasets demands substantial resources.
- Few-shot semantic segmentation learns to segment target classes (novel classes) in the *query* image using only a few pixel-wise annotated *support* images, enabling segmentation models trained on *base* classes to generalize to *novel* classes.



- Nico Catalano, and Matteo Matteucci. “Few Shot Semantic Segmentation: a review of methodologies, benchmarks, and open challenges”. arXiv preprint arXiv:2304.05832 (2023).

# Few-shot Image Semantic Segmentation

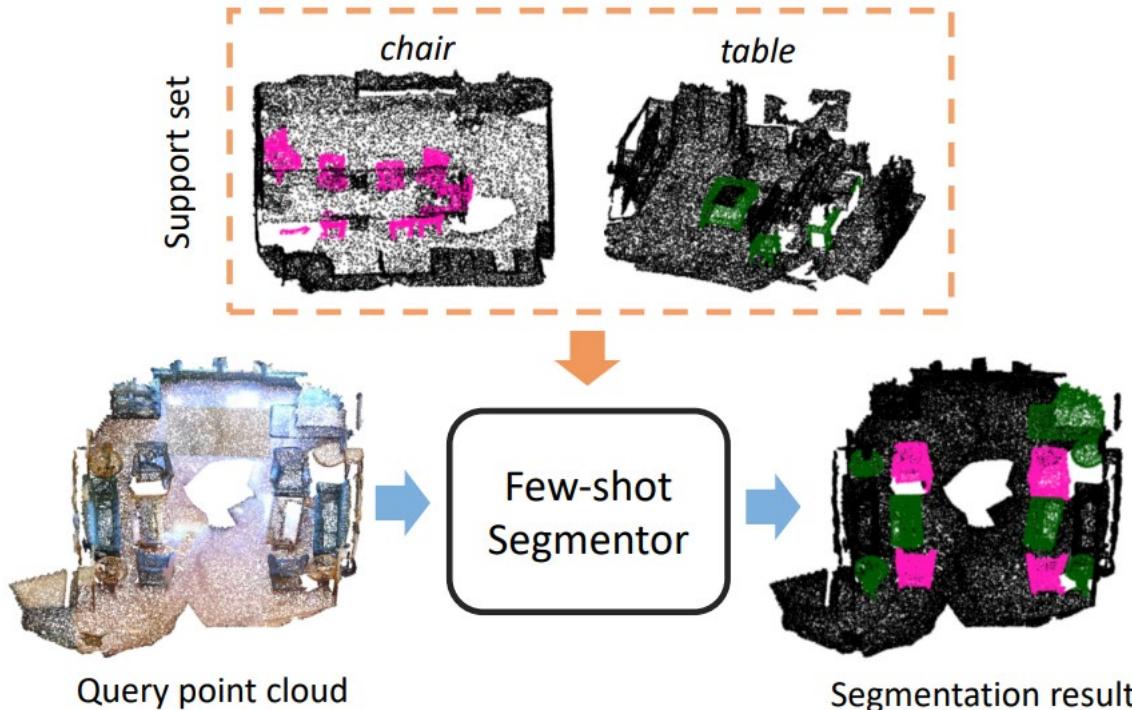
- A shared feature extractor gets a feature volume from both the support set and query images. The **Masked Average Pooling (MAP)** module takes the feature volume from the support set and masks its ground truth with the Hadamard product  $\odot$  to compute the class prototype.
- The prediction mask  $\hat{M}_q$  is calculated as a **metric** between the vector at each spatial location in the query feature volume with the class prototype.



- Nico Catalano, and Matteo Matteucci. “Few Shot Semantic Segmentation: a review of methodologies, benchmarks, and open challenges”. arXiv preprint arXiv:2304.05832 (2023).

# Few-shot 3D Point Cloud Semantic Segmentation

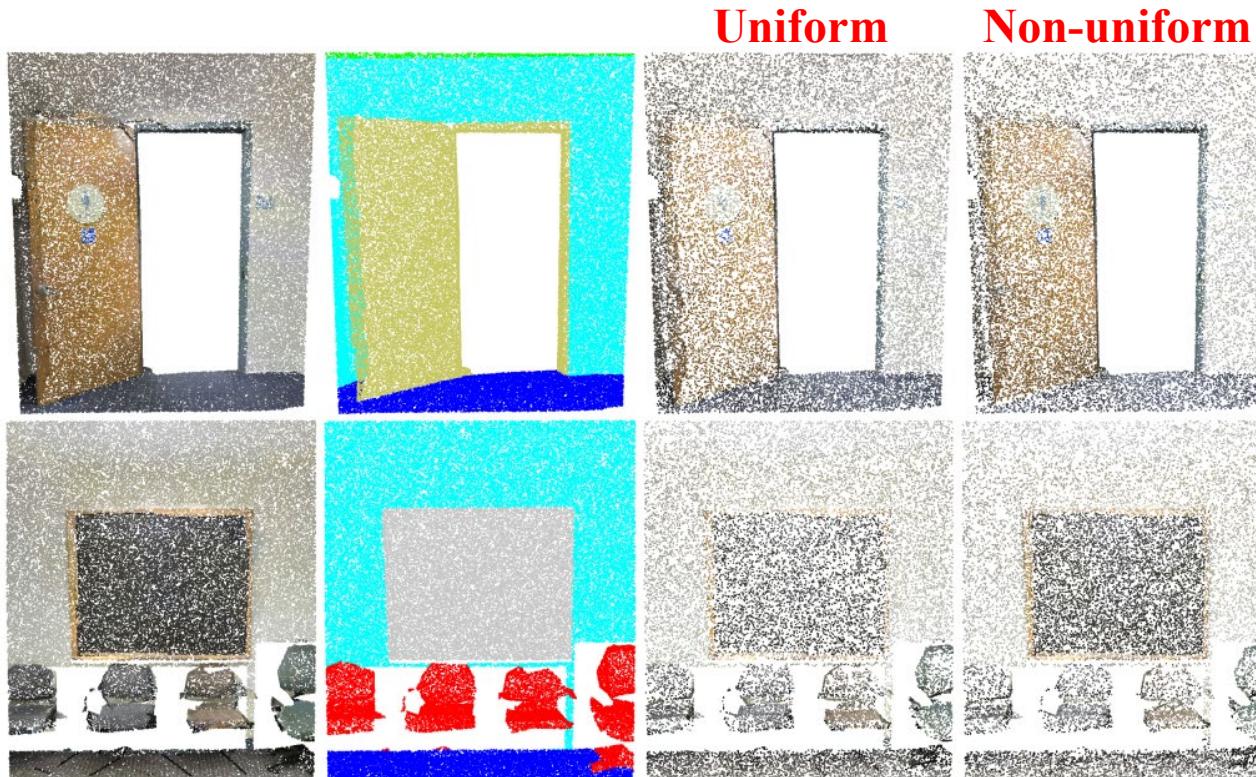
- Few-shot point cloud semantic segmentation (**FS-PCS**) learns to segment **target classes** in the *query* point cloud given a few annotated *support* point clouds. This figure illustrates an example with the 2-way 1-shot setting, which means that we have two target classes (chair and table) and one support point cloud for each class.



- Na Zhao, Tat-Seng Chua, and Gim Hee Lee. “Few-shot 3d point cloud semantic segmentation”. In Proceedings of the IEEE/CVF CVPR, pp. 8873-8882. 2021.

# Issue 1: Foreground Leakage

- The point sampling process in previous FS-PCS is non-uniform, favoring more points in the foreground than in the background. This leads to foreground leakage, a noticeable density bias toward foreground classes.



From left to right: (1) The original point cloud; (2) Ground truth of all categories; (3) Our corrected input with 20,480 points in a uniform distribution; (4) Input with 20,480 points in a biased distribution.

# Issue 1: Foreground Leakage

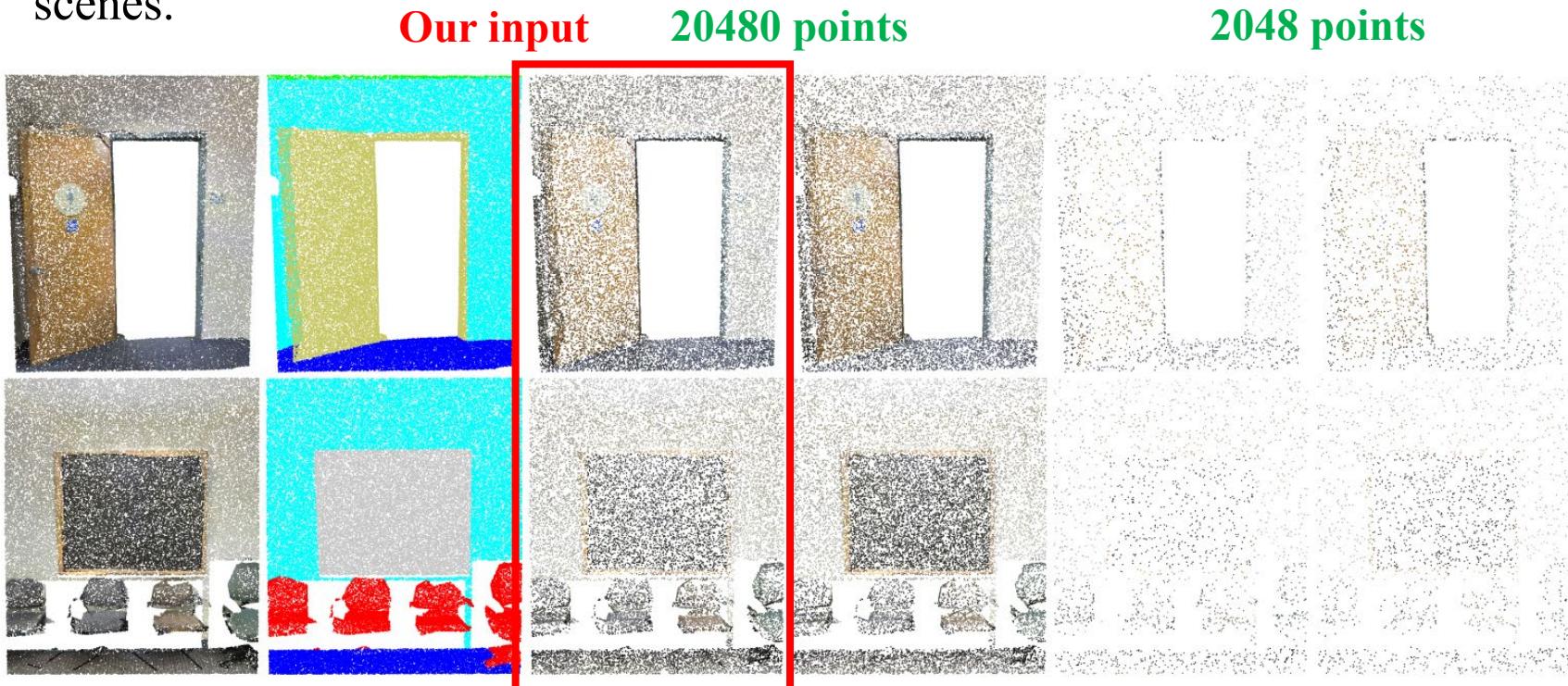
- The current non-uniform point sampling leads to a noticeable **point density disparity** between foreground and background, which induces models to segment foreground classes by identifying **denser** regions, instead of learning semantic knowledge transfer from support to query.
- Addressing this issue results in a significant performance drop in existing methods.

|        | Methods      | 1-shot (S3DIS) |       |       | 5-shot (S3DIS) |       |       | 1-shot (ScanNet) |       |       | 5-shot (ScanNet) |       |       |
|--------|--------------|----------------|-------|-------|----------------|-------|-------|------------------|-------|-------|------------------|-------|-------|
|        |              | $S^0$          | $S^1$ | mean  | $S^0$          | $S^1$ | mean  | $S^0$            | $S^1$ | mean  | $S^0$            | $S^1$ | mean  |
| w/ FG  | AttMPTI [56] | 64.89          | 66.15 | 65.52 | 76.56          | 83.08 | 79.82 | 62.14            | 58.65 | 60.39 | 68.79            | 68.66 | 68.73 |
|        | QGE [29]     | 74.05          | 73.61 | 73.83 | 74.65          | 83.21 | 78.93 | 63.50            | 57.61 | 60.56 | 70.72            | 65.68 | 68.20 |
|        | QGPA [11]    | 62.72          | 61.95 | 62.33 | 76.30          | 87.29 | 81.80 | 56.47            | 51.72 | 54.10 | 81.57            | 72.75 | 77.16 |
| w/o FG | AttMPTI [56] | 41.56          | 41.27 | 41.41 | 50.55          | 46.13 | 48.34 | 33.36            | 31.81 | 32.58 | 37.95            | 36.30 | 37.12 |
|        | QGE [29]     | 46.27          | 47.76 | 47.02 | 47.74          | 59.77 | 53.76 | 37.72            | 34.64 | 36.18 | 48.73            | 39.95 | 44.34 |
|        | QGPA [11]    | 35.62          | 41.13 | 38.38 | 43.54          | 47.50 | 45.52 | 40.03            | 35.54 | 37.78 | 46.17            | 42.24 | 44.20 |

**Comparisons in the mIoU metric between *with* foreground leakage (*w/ FG*) and *without* foreground leakage (*w/o FG*) for existing methods.** The results are for 1-way segmentation setting.  $S^0/S^1$  refers to the  $i$ -th split for inference.

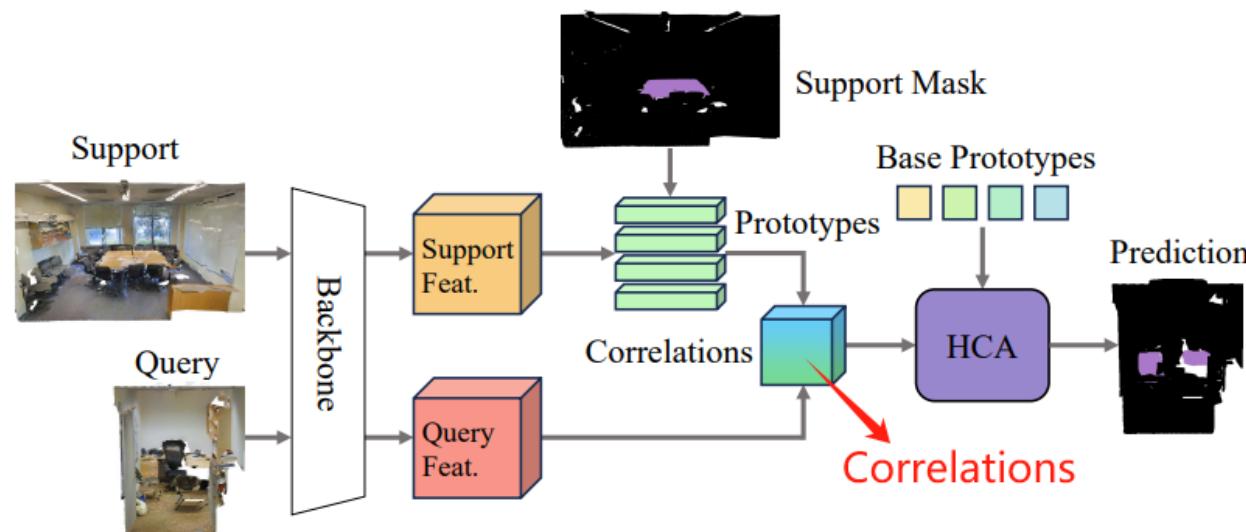
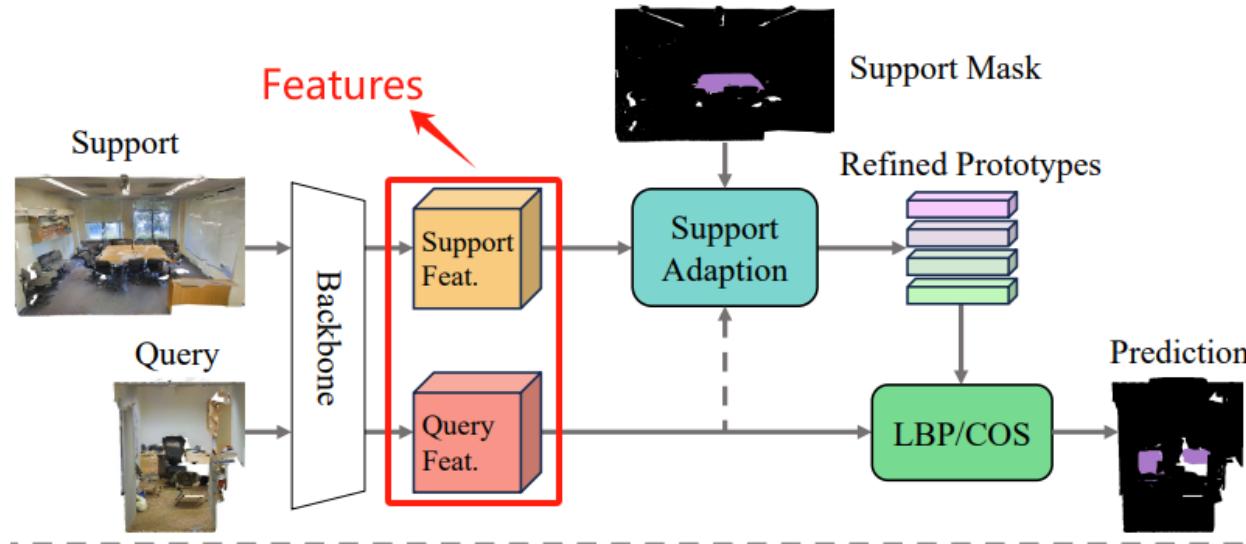
## Issue 2: Sparse Point Distribution

- The current FSPCS input is constrained to only 2,048 points.
- These sparsely distributed, semantically limited inputs introduce significant ambiguities, hindering the model's capacity to exploit semantics in the scenes.



From left to right: (1) The original point cloud; (2) Ground truth of all categories; (3) Our corrected input with 20,480 points in a uniform distribution; (4) Input with 20,480 points in a biased distribution; (5) Input with 2,048 points in a uniform distribution; (6) Input with 2,048 points in a biased distribution, as adopted by previous works.

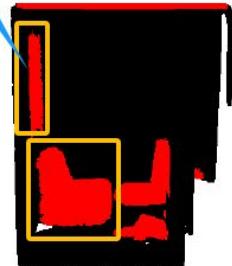
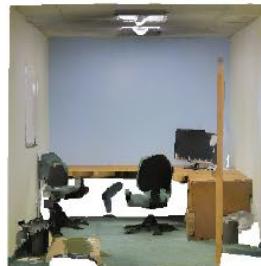
# Motivation 1: Feature Optimization vs. Correlation Optimization



*Top:* Previous feature optimization; *Bottom:* Our correlation optimization

## Motivation 2: Base Susceptibility Problem

- Within the meta-learning framework, models undergo training on *seen/base* classes and are evaluated on *unseen/novel* classes.
- These models tend to be susceptible to the base classes within test scenes, thereby hindering the accurate segmentation of novel classes.



Support

Support Mask

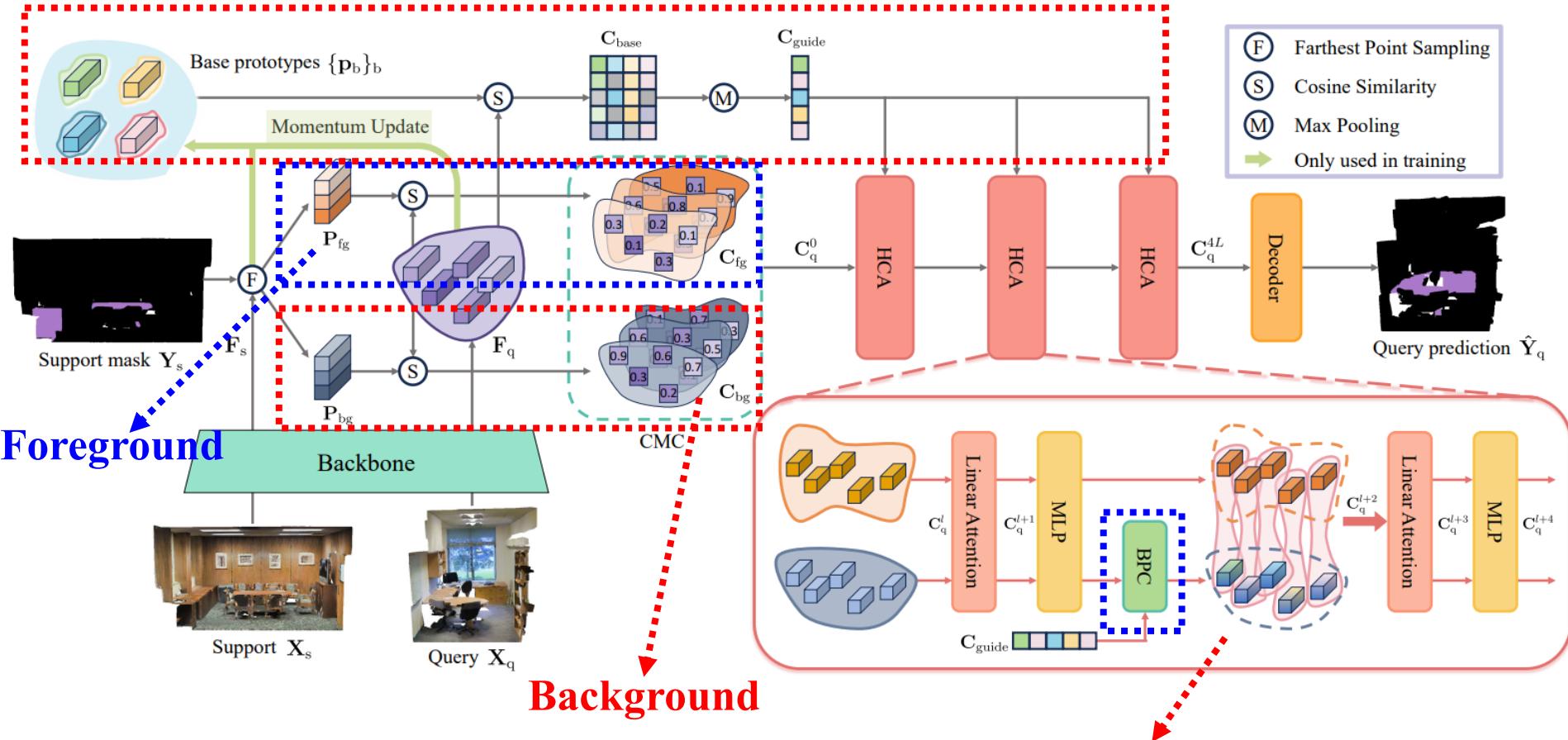
Query

Ground Truth

Prediction

# Correlation Optimization Segmentation (COSeg)

## Base Prototypes Calibration (BPC)



HCA: Hyper Correlation Augmentation, a carefully designed module for query-support correlation optimization. Please see the paper for details.

- 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”. **IEEE CVPR, 2024**.

# Class-specific Multi-prototypical Correlation (CMC)

- Foreground prototypes  $\mathbf{P}_{\text{fg}}$  and background prototypes  $\mathbf{P}_{\text{bg}}$  are obtained through two steps: sample seeds in the coordinate space based on farthest point sampling, and then conduct point-to-seed clustering as follows:

$$\mathbf{P}_{\text{fg}} = \mathcal{F}_{\text{clus}}(\mathbf{F}_s \odot \mathbf{Y}_s, \mathbf{S}_{\text{fg}}), \quad \mathbf{S}_{\text{fg}} = \mathcal{F}_{\text{fps}}(\mathbf{L}_s \odot \mathbf{Y}_s),$$

$$\mathbf{P}_{\text{bg}} = \mathcal{F}_{\text{clus}}(\mathbf{F}_s \odot \tilde{\mathbf{Y}}_s, \mathbf{S}_{\text{bg}}), \quad \mathbf{S}_{\text{bg}} = \mathcal{F}_{\text{fps}}(\mathbf{L}_s \odot \tilde{\mathbf{Y}}_s)$$

- We compute the cosine similarities of query points with respect to  $\mathbf{P}_{\text{fg}}$  and  $\mathbf{P}_{\text{bg}}$ , and obtain the correlations:

$$\mathbf{C}_{\text{fg}} = \frac{\mathbf{F}_q \cdot \mathbf{P}_{\text{fg}}^\top}{\|\mathbf{F}_q\| \|\mathbf{P}_{\text{fg}}^\top\|}, \quad \mathbf{C}_{\text{bg}} = \frac{\mathbf{F}_q \cdot \mathbf{P}_{\text{bg}}^\top}{\|\mathbf{F}_q\| \|\mathbf{P}_{\text{bg}}^\top\|}$$

- We concatenate  $\mathbf{C}_{\text{fg}}$  and  $\mathbf{C}_{\text{bg}}$  along the second dimension and project the last dimension back to  $D$  using an MLP, as follows:

$$\mathbf{C}_q^0 = \mathcal{F}_{\text{mlp}}(\mathbf{C}_{\text{fg}} \oplus \mathbf{C}_{\text{bg}}) \in \mathbb{R}^{N_Q \times N_C \times D}$$

- 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”. **IEEE CVPR, 2024**.

# Hyper Correlation Augmentation (HCA)

- We permute  $\mathbf{C}_q^l$  with the class dimension as its first dimension and then compute linear attention across points:

$$\mathbf{C}_q^{l+1} = \mathcal{F}_{\text{lnatt}}(\mathcal{T}(\mathbf{C}_q^l)) \in \mathbb{R}^{N_C \times N_Q \times D}$$

- Following the attention layer, an MLP is applied:

$$\mathbf{C}_q^{l+2} = \mathcal{F}_{\text{mlp}}(\mathbf{C}_q^{l+1}) \in \mathbb{R}^{N_C \times N_Q \times D}$$

- We rearrange the dimensions and apply linear attention, given by:

$$\mathbf{C}_q^{l+3} = \mathcal{F}_{\text{lnatt}}(\mathcal{T}(\mathbf{C}_q^{l+2})) \in \mathbb{R}^{N_Q \times N_C \times D}$$

- 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”. **IEEE CVPR, 2024**.

# Base Prototypes Calibration (BPC)

- During meta learning, we calculate the Masked Average Pooling (MAP) for each base class present in the current point clouds:

$$\begin{aligned}\mathbf{p}'_b &= \mathcal{F}_{\text{pool}}(\mathbf{F}_{s/q} \odot \mathbf{Y}_{s/q}^b) \in \mathbb{R}^{1 \times D}, \\ \mathbf{p}_b &\leftarrow \mu \mathbf{p}_b + (1 - \mu) \mathbf{p}'_b\end{aligned}$$

- We calculate the base correlations  $\mathbf{C}_{\text{base}}$  between the query and base prototypes:

$$\mathbf{C}_{\text{base}} = \frac{\mathbf{F}_q \cdot \mathcal{I}(\{\mathbf{p}_b\}_{b=1}^{N_b})^\top}{\|\mathbf{F}_q\| \left\| \mathcal{I}(\{\mathbf{p}_b\}_{b=1}^{N_b})^\top \right\|} \in \mathbb{R}^{N_Q \times N_b}$$

- The background correlations are calibrated by  $\mathbf{C}_{\text{guide}}$  before interacting with foreground correlations:

$$\mathbf{C}_{\text{guide}} = \mathcal{F}_{\text{max}}(\mathbf{C}_{\text{base}}) \in \mathbb{R}^{N_Q}$$

$$\mathbf{C}_q^{l+2}[1, \cdot, \cdot] = \mathcal{F}_{\text{fc}}(\mathbf{C}_q^{l+2}[1, \cdot, \cdot] \oplus \mathcal{D}(\mathbf{C}_{\text{guide}}))$$

- 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”. **IEEE CVPR, 2024**.

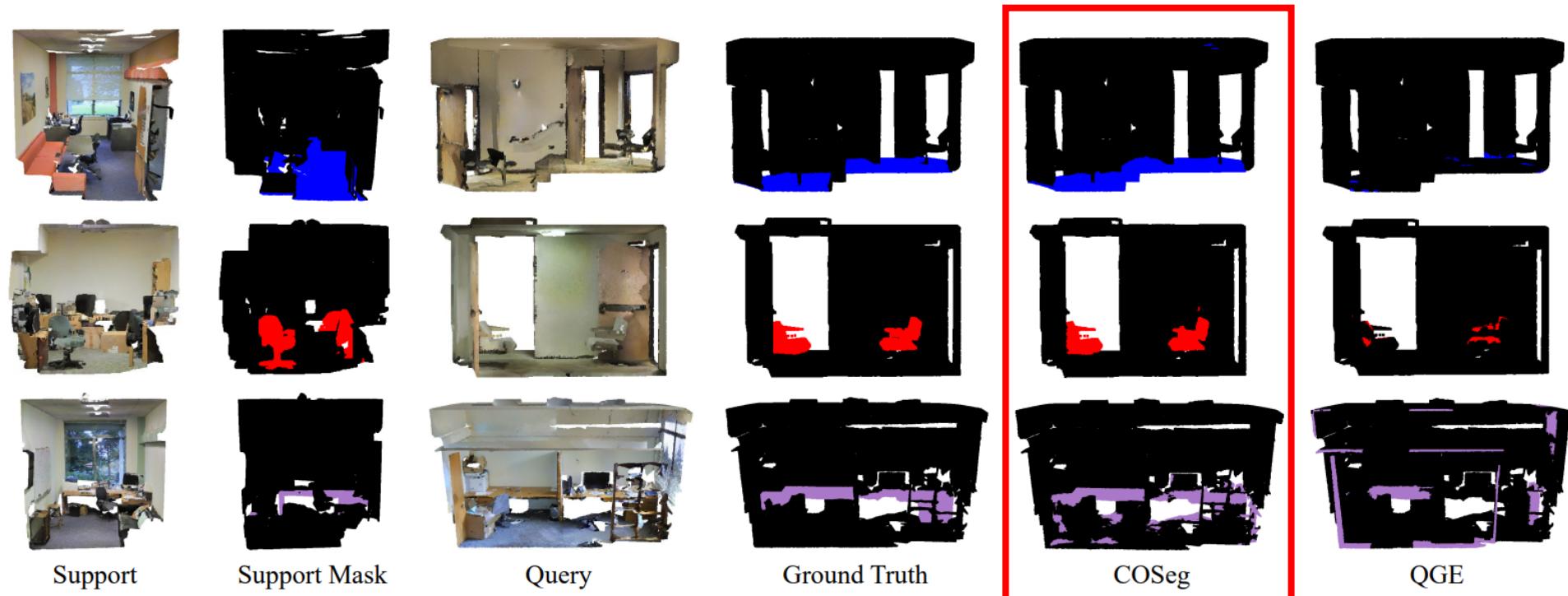
# Quantitative Comparison

|             | Methods      | 1-way 1-shot |              |              | 1-way 5-shot |              |              | 2-way 1-shot |              |              | 2-way 5-shot |              |              |
|-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|             |              | $S^0$        | $S^1$        | mean         |
| S3DIS [1]   | AttMPTI [56] | 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 [29]     | 41.69        | 39.09        | 40.39        | 50.59        | 46.41        | 48.50        | 33.45        | 30.95        | 32.20        | 40.53        | 36.13        | 38.33        |
|             | QGPA [11]    | 35.50        | 35.83        | 35.67        | 38.07        | 39.70        | 38.89        | 25.52        | 26.26        | 25.89        | 30.22        | 32.41        | 31.32        |
|             | COSeg (ours) | <b>46.31</b> | <b>48.10</b> | <b>47.21</b> | <b>51.40</b> | <b>48.68</b> | <b>50.04</b> | <b>37.44</b> | <b>36.45</b> | <b>36.95</b> | <b>42.27</b> | <b>38.45</b> | <b>40.36</b> |
| ScanNet [7] | AttMPTI [56] | 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 [29]     | 37.38        | 33.02        | 35.20        | 45.08        | 41.89        | 43.49        | 26.85        | 25.17        | 26.01        | 28.35        | 31.49        | 29.92        |
|             | QGPA [11]    | 34.57        | 33.37        | 33.97        | 41.22        | 38.65        | 39.94        | 21.86        | 21.47        | 21.67        | 30.67        | 27.69        | 29.18        |
|             | COSeg (ours) | <b>41.73</b> | <b>41.82</b> | <b>41.78</b> | <b>48.31</b> | <b>44.11</b> | <b>46.21</b> | <b>28.72</b> | <b>28.83</b> | <b>28.78</b> | <b>35.97</b> | <b>33.39</b> | <b>34.68</b> |

Comparisons in the mIoU metric between our method and baselines in the new FS-PCS setting.

- 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”. **IEEE CVPR, 2024**.

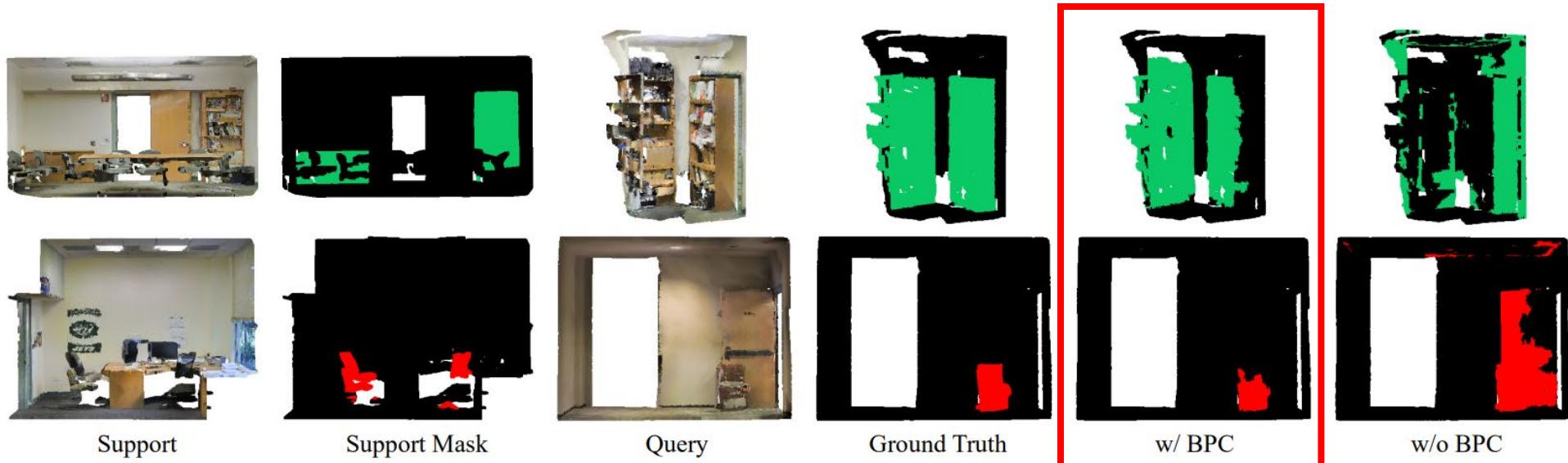
# Qualitative Comparison



Qualitative comparisons between our proposed model COSeg and QGE (SOTA method). Each row, from top to bottom, represents the 1-way 1-shot task with the target category as floor (blue), chair (red), and table (purple), respectively.

- 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”. **IEEE CVPR, 2024**.

# Qualitative Comparison



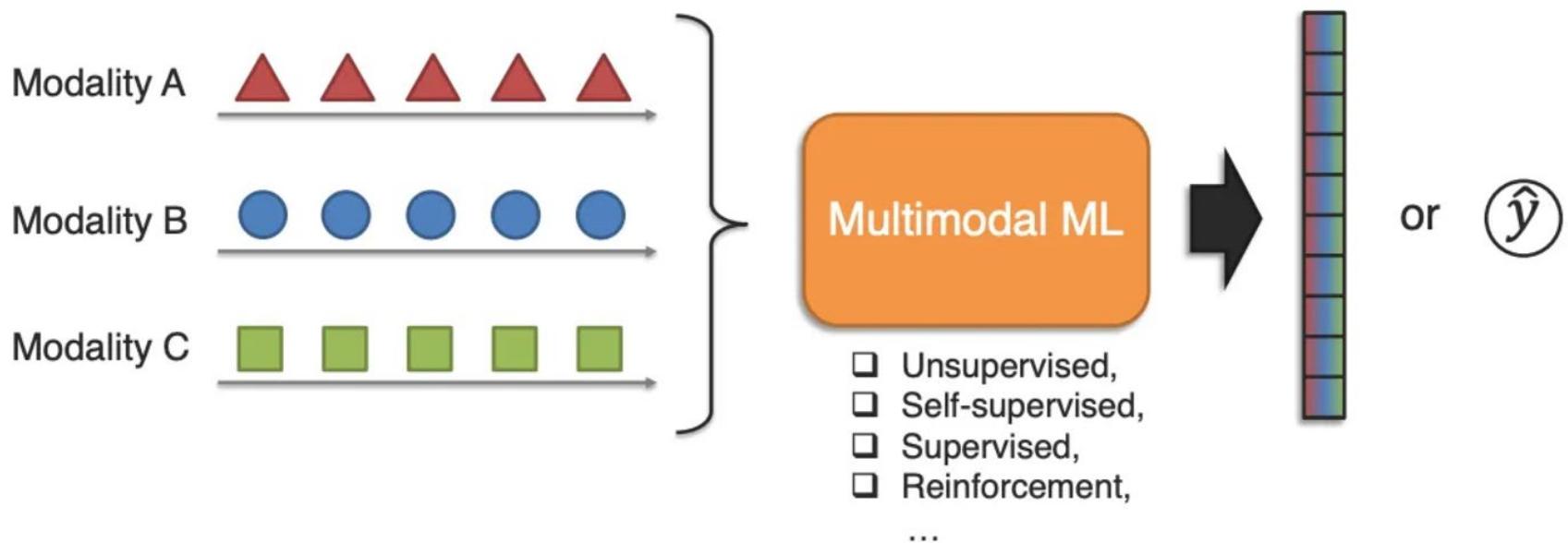
Visual comparisons between our models *with* BPC (*w/ BPC*) and *without* BPC (*w/o BPC*). Each row corresponds to the 1-way 1-shot task targeting bookcase (green) and chair (red), respectively, arranged from top to bottom.

BPC: Base Prototypes Calibration

**Code:** <https://github.com/ZhaochongAn/COSeg>

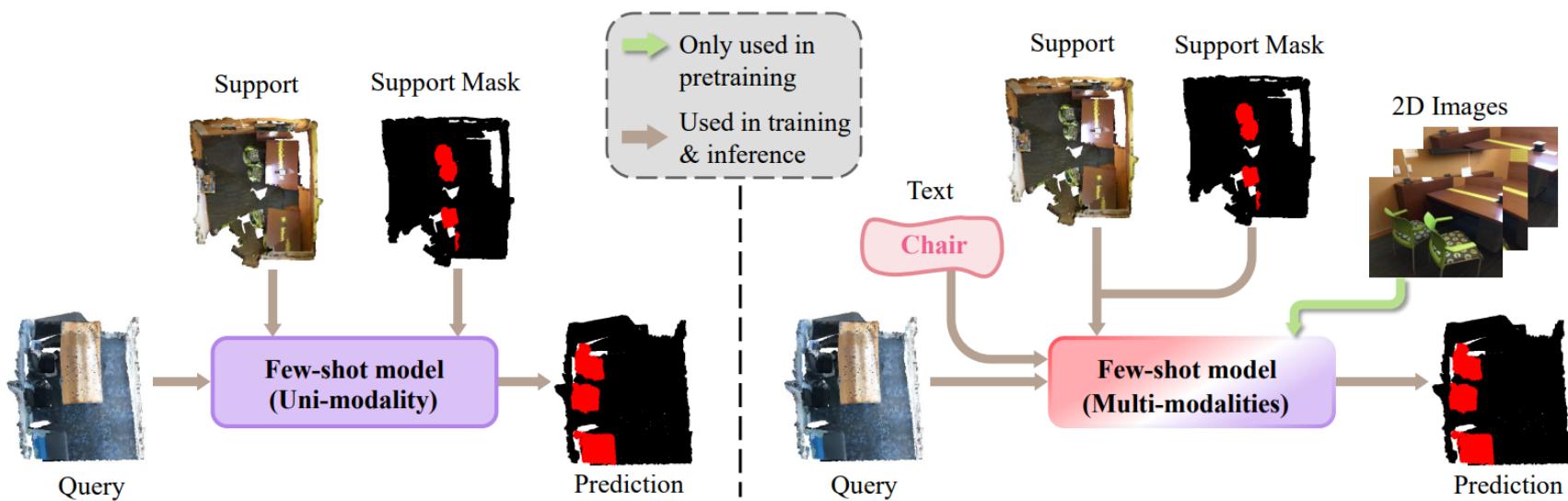
# Multimodal Learning

- Multimodal learning is a type of deep learning that integrates and processes multiple types of data, referred to as modalities, such as text, audio, images, or video.
- Large multimodal models, such as Google Gemini and GPT-4o, have become increasingly popular since 2023, enabling increased versatility and a broader understanding of real-world phenomena.



# Unimodal FS-PCS vs. Multimodal FS-PCS

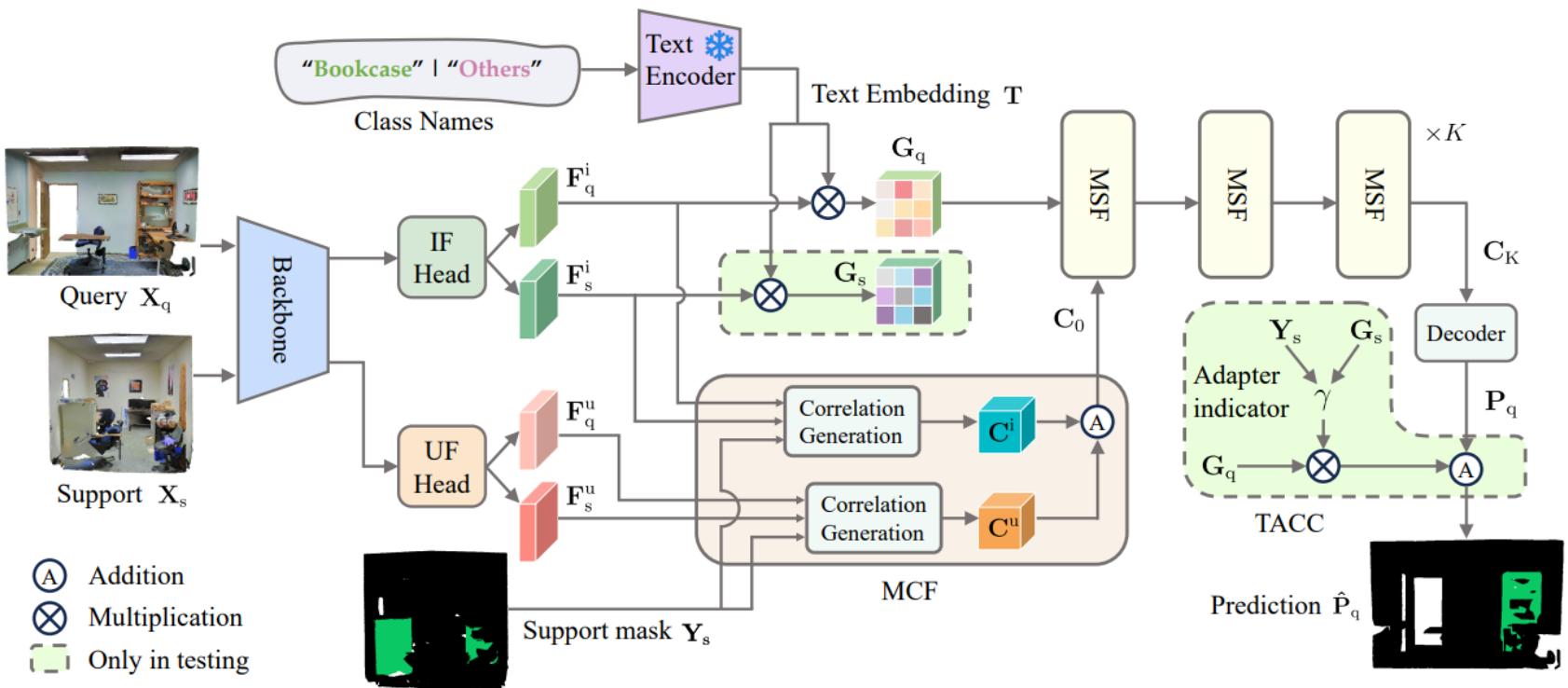
- Previous FS-PCS methods (left) only make use of point clouds as unimodal input. In contrast, our proposed model (right) utilizes **cost-free** multimodal information to improve FS-PCS by considering the textual modality of class names (*explicit*) and learning the simulated features of the 2D modality (*implicit*). During meta-learning and inference, the 2D modality is not needed.



- 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”. **ICLR 2025**.

# Overall Architecture of Multimodal FS-PCS

- Given support and query point clouds, we first generate intermodal features  $F_{s/q}^i$  from the IF head and unimodal features  $F_{s/q}^u$  from the UF head. These features are then forwarded to the MCF module to generate initial multimodal correlations  $C_0$ .



- 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". **ICLR 2025**.

# Feature Extractors

- The IF head extracts intermodal features that are aligned with 2D visual features by exploiting the 2D modality, while the UF head focuses solely on the 3D point cloud modality.

$$\mathbf{F}_s^i = \mathcal{H}_{\text{IF}}(\mathbf{F}_s) \in \mathbb{R}^{N_S \times D_t}, \quad \mathbf{F}_s^u = \mathcal{H}_{\text{UF}}(\mathbf{F}_s) \in \mathbb{R}^{N_S \times D},$$
$$\mathbf{F}_q^i = \mathcal{H}_{\text{IF}}(\mathbf{F}_q) \in \mathbb{R}^{N_Q \times D_t}, \quad \mathbf{F}_q^u = \mathcal{H}_{\text{UF}}(\mathbf{F}_q) \in \mathbb{R}^{N_Q \times D}.$$

- In the pretraining, we employ a cosine similarity loss to minimize the distance between 3D point intermodal features and corresponding 2D pixel features. Then, we fix the backbone and IF head during meta-learning.
- We compute embeddings for the “background” and target classes using the LSeg text encoder, denoted as  $\mathbf{T} = \{\mathbf{t}_0, \dots, \mathbf{t}_N\} \in \mathbb{R}^{N_C \times D_t}$
- 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”. **ICLR 2025**.

# Multimodal Correlation Fusion (MCF)

- Prototypes are generated from the annotated support points for both intermodal and unimodal features. The correlations between the query points and these prototypes:

$$\mathbf{C}^i = \frac{\mathbf{F}_q^i \cdot \mathbf{P}_{proto}^{i\top}}{\|\mathbf{F}_q^i\| \|\mathbf{P}_{proto}^{i\top}\|}, \quad \mathbf{C}^u = \frac{\mathbf{F}_q^u \cdot \mathbf{P}_{proto}^{u\top}}{\|\mathbf{F}_q^u\| \|\mathbf{P}_{proto}^{u\top}\|}$$

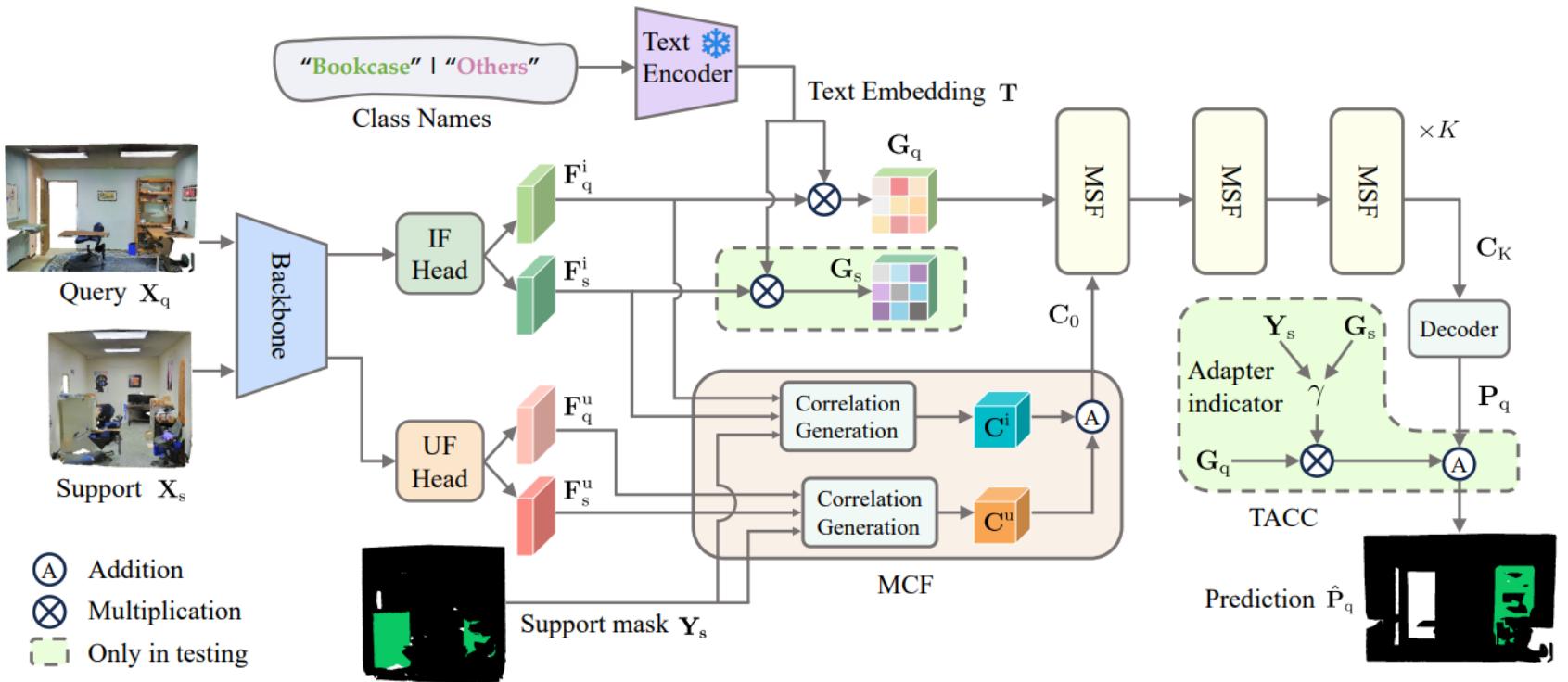
- The MCF module transforms these correlations using two linear layers and then combines them to obtain the aggregated multimodal correlation  $\mathbf{C}_0$ , as follows:

$$\mathbf{C}_0 = \mathcal{F}_{lin}(\mathbf{C}^i) + \mathcal{F}_{lin}(\mathbf{C}^u), \quad \mathbf{C}_0 \in \mathbb{R}^{N_Q \times N_C \times D}$$

- 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”. **ICLR 2025**.

# Overall Architecture of Multimodal FS-PCS

- For exploiting the alignment between intermodal features  $F_q^i$  and text embeddings  $T$ , we use their affinity  $G_q$  as the informative textual semantic guidance to refine the multimodal correlations in the MSF modules. Finally, we propose the TACC, a parameter-free module that adaptively calibrates predictions during test time to effectively mitigate the base bias issue.



- 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". **ICLR 2025**.

# Multimodal Semantic Fusion (MSF)

- We first compute the similarity between the query intermodal features and text embeddings to generate semantic guidance:

$$\mathbf{G}_q = \mathbf{F}_q^i \cdot \mathbf{T}^\intercal$$

- Point-category weights to consider varying importance between visual and textual modalities are dynamically computed as follows:

$$\mathbf{W}_q = \mathcal{F}_{\text{mlp}}(\mathcal{F}_{\text{expand}}(\mathbf{G}_q) \oplus \mathbf{C}_k), \quad \mathbf{W}_q \in \mathbb{R}^{N_Q \times N_C \times 1}$$

- The semantic guidance  $\mathbf{G}_q$ , weighted by  $\mathbf{W}_q$ , is aggregated into the correlation input  $\mathbf{C}_k$ :

$$\begin{aligned}\mathbf{C}'_k &= \mathbf{G}_q \odot \mathbf{W}_q + \mathbf{C}_k, \\ \mathbf{C}_{k+1} &= \mathcal{F}_{\text{mlp}}(\mathcal{F}_{\text{attention}}(\mathbf{C}'_k))\end{aligned}$$

- 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”. **ICLR 2025**.

# Test-time Adaptive Cross-modal Calibration (TACC)

- We propose an adaptive combination of the semantic guidance  $\mathbf{G}_q$  and the prediction  $\mathbf{P}_q$  through an adaptive indicator  $\gamma$ :

$$\hat{\mathbf{P}}_q = \gamma \mathbf{G}_q + \mathbf{P}_q$$

- Using the support intermodal features  $F$  and the text embeddings  $T$ , we compute  $\mathbf{G}_s$ , which is then used to generate the predicted labels  $\mathbf{P}_s$ . With the available support labels  $\mathbf{Y}_s$  in each episode, the quality of  $\mathbf{G}_s/\mathbf{G}_q$  is quantified by comparing the predicted labels  $\mathbf{P}_s$  to  $\mathbf{Y}_s$  using the Intersection-over-Union (IoU) score.

$$\gamma = \frac{\sum_i \mathbf{1}_{\{\mathbf{P}_s(i)=1 \wedge \mathbf{Y}_s(i)=1\}}}{\sum_i \mathbf{1}_{\{\mathbf{P}_s(i)=1 \vee \mathbf{Y}_s(i)=1\}}}, \quad \mathbf{P}_s[i] = \arg \max(\mathbf{G}_s[i, :]), \quad \mathbf{G}_s = \mathbf{F}_s^i \cdot \mathbf{T}^\top$$

- 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”. **ICLR 2025**.

# Quantitative Comparison

| Methods                              | 1-way 1-shot |              |                    | 1-way 5-shot |              |                    | 2-way 1-shot |              |                    | 2-way 5-shot |              |                    |
|--------------------------------------|--------------|--------------|--------------------|--------------|--------------|--------------------|--------------|--------------|--------------------|--------------|--------------|--------------------|
|                                      | $S^0$        | $S^1$        | Mean               |
| 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              |
| QGPA (He et al., 2023)               | 35.50        | 35.83        | 35.67              | 38.07        | 39.70        | 38.89              | 25.52        | 26.26        | 25.89              | 30.22        | 32.41        | 31.32              |
| COSeg (An et al., 2024)              | 46.31        | 48.10        | 47.21              | 51.40        | 48.68        | 50.04              | 37.44        | 36.45        | 36.95              | 42.27        | 38.45        | 40.36              |
| COSeg <sup>†</sup> (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 (ours)                        | <b>49.84</b> | <b>54.33</b> | <b>52.09(+4.3)</b> | <b>51.95</b> | <b>56.46</b> | <b>54.21(+3.8)</b> | <b>41.98</b> | <b>46.61</b> | <b>44.30(+6.2)</b> | <b>46.02</b> | <b>54.29</b> | <b>50.16(+8.7)</b> |

Table 1: **Quantitative comparison with previous methods in mIoU (%) on the S3DIS dataset.** There are four few-shot settings: 1/2-way 1/5-shot.  $S^0/S^1$  refers to using the split  $i$  for evaluation, and ‘Mean’ represents the average mIoU on both splits. The best results are highlighted in **bold**.

| Methods                              | 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               |
| 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              |
| QGPA (He et al., 2023)               | 34.57        | 33.37        | 33.97              | 41.22        | 38.65        | 39.94              | 21.86        | 21.47        | 21.67               | 30.67        | 27.69        | 29.18              |
| COSeg (An et al., 2024)              | 41.73        | 41.82        | 41.78              | 48.31        | 44.11        | 46.21              | 28.72        | 28.83        | 28.78               | 35.97        | 33.39        | 34.68              |
| COSeg <sup>†</sup> (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 (ours)                        | <b>46.08</b> | <b>43.37</b> | <b>44.73(+2.7)</b> | <b>54.66</b> | <b>45.48</b> | <b>50.07(+3.5)</b> | <b>43.99</b> | <b>34.43</b> | <b>39.21(+10.2)</b> | <b>48.86</b> | <b>39.32</b> | <b>44.09(+8.6)</b> |

Table 2: **Quantitative comparison with previous methods in mIoU (%) on the ScanNet dataset.**

- 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”. **ICLR 2025**.

# Qualitative Comparison

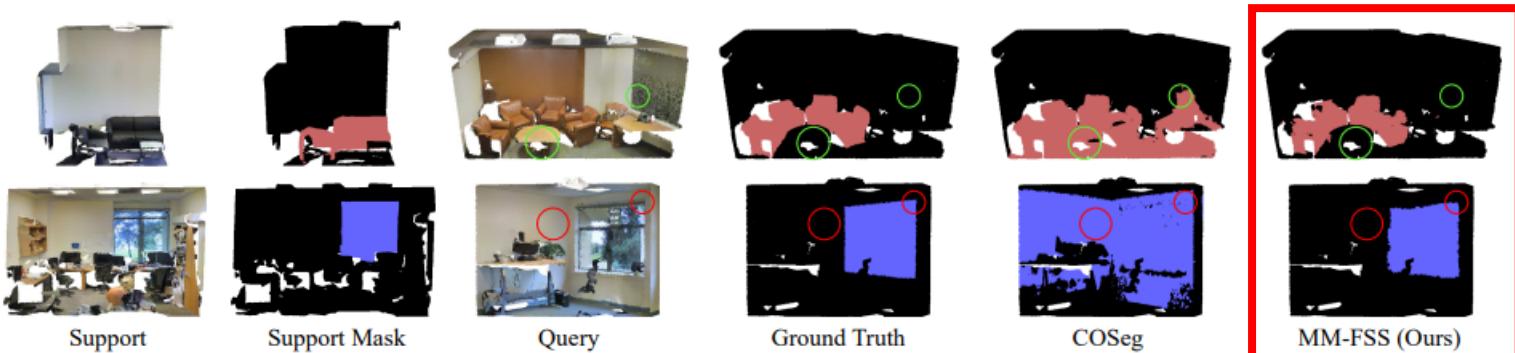


Figure 3: Qualitative comparison between COSeg and our proposed MM-FSS in the 1-way 1-shot setting on the S3DIS dataset. The target classes in the first and second rows are sofa and window, respectively. Important areas are marked with circles.

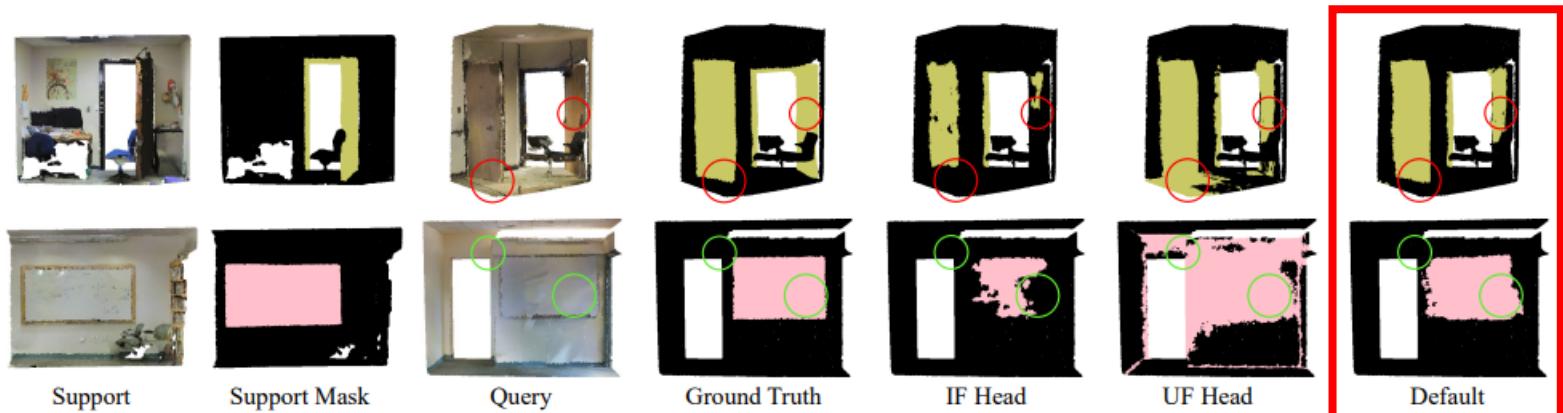
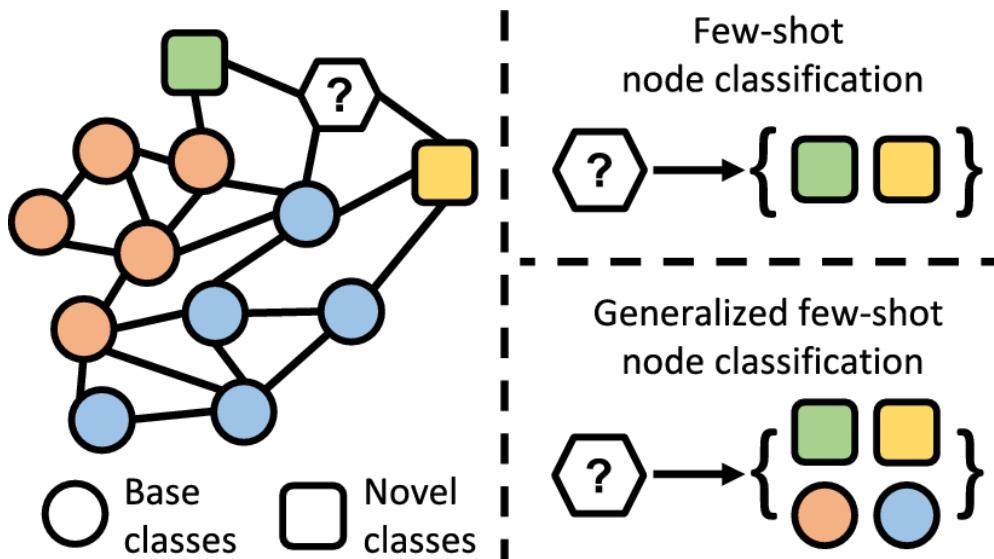


Figure 4: Qualitative comparison of predictions from each head and our final prediction using TACC (Default) in the 1-way 1-shot setting on the S3DIS dataset. The target classes in the first and second rows are door and board, respectively.

- 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”. ICLR 2025.

# Few-shot Seg. vs. Generalized Few-shot Seg.

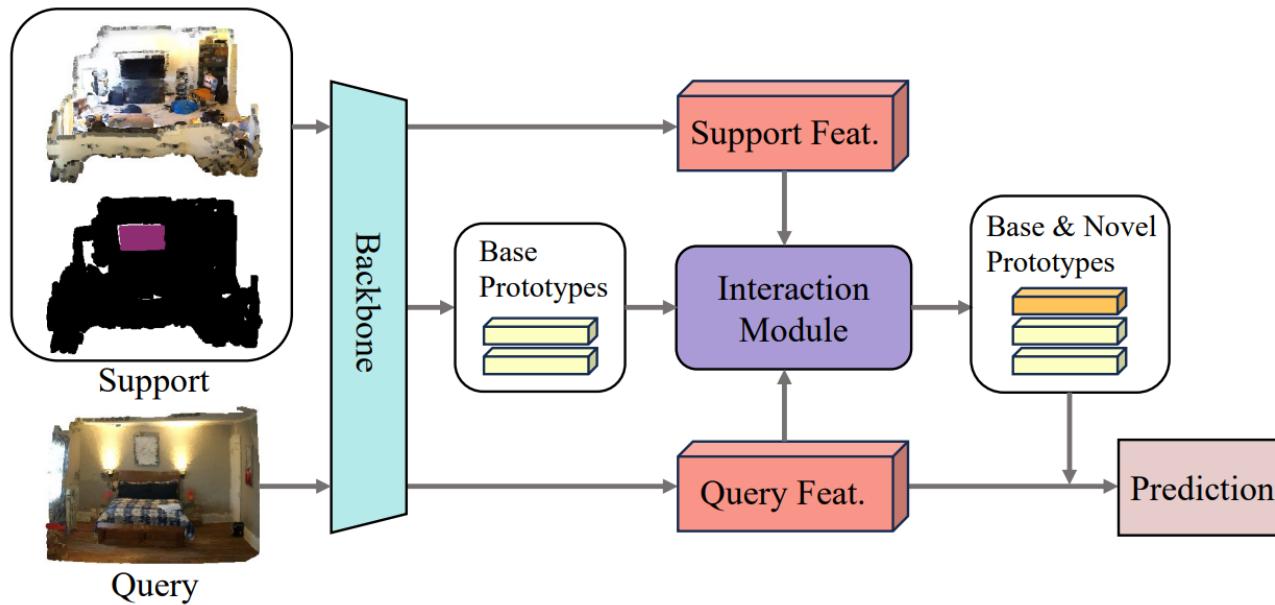
- Few-shot segmentation requires additional support samples for each novel class at inference and **only predicts novel classes**, ignoring base classes.
- Generalized Few-shot segmentation directly segments **both base and novel classes** after few-shot adaptation, making it more practical for real-world applications.
- **GFS-PCS: Generalized Few-shot 3D Point Cloud Segmentation**



- Zhe Xu, Kaize Ding, Yu-Xiong Wang, Huan Liu, and Hanghang Tong. “Generalized few-shot node classification: toward an uncertainty-based solution”. **Knowledge and Information Systems** 2024.

# Challenge in GFS-PCS

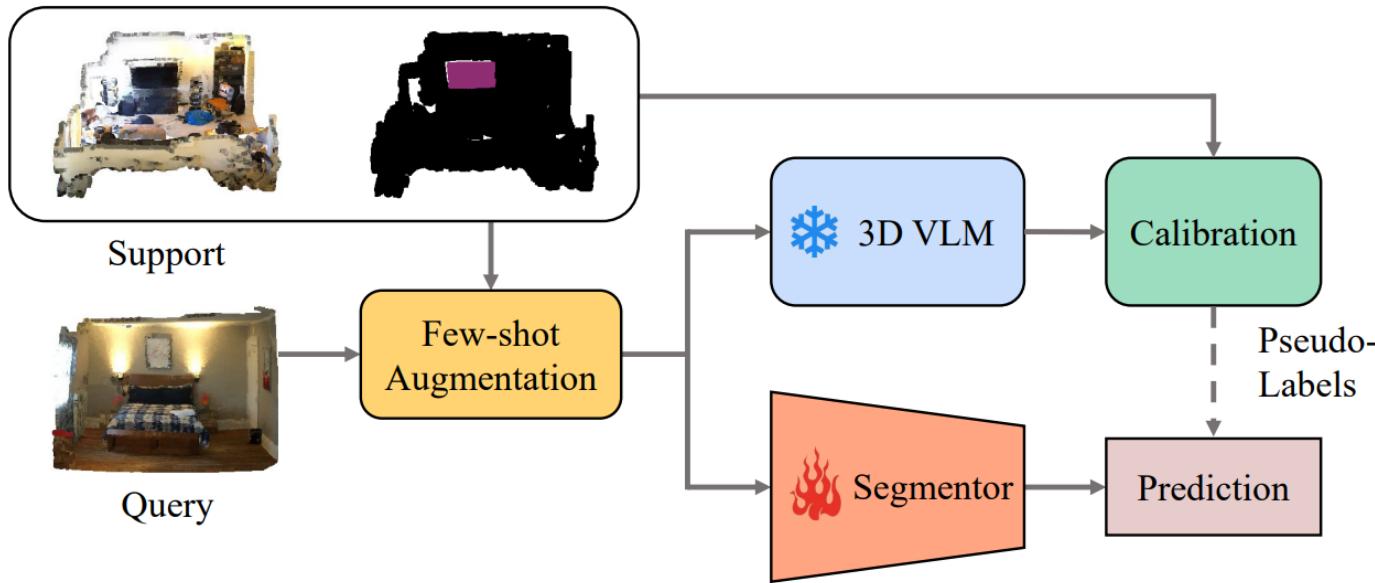
- Prior work primarily enhances prototypes through interaction modules that integrate support/query features, making predictions based on refined prototypes. However, they are **limited by the sparse knowledge from few-shot samples.**



- 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”. **IEEE CVPR, 2025**.

# GFS-PCS with 3D Vision-Language Models (3D VLMs)

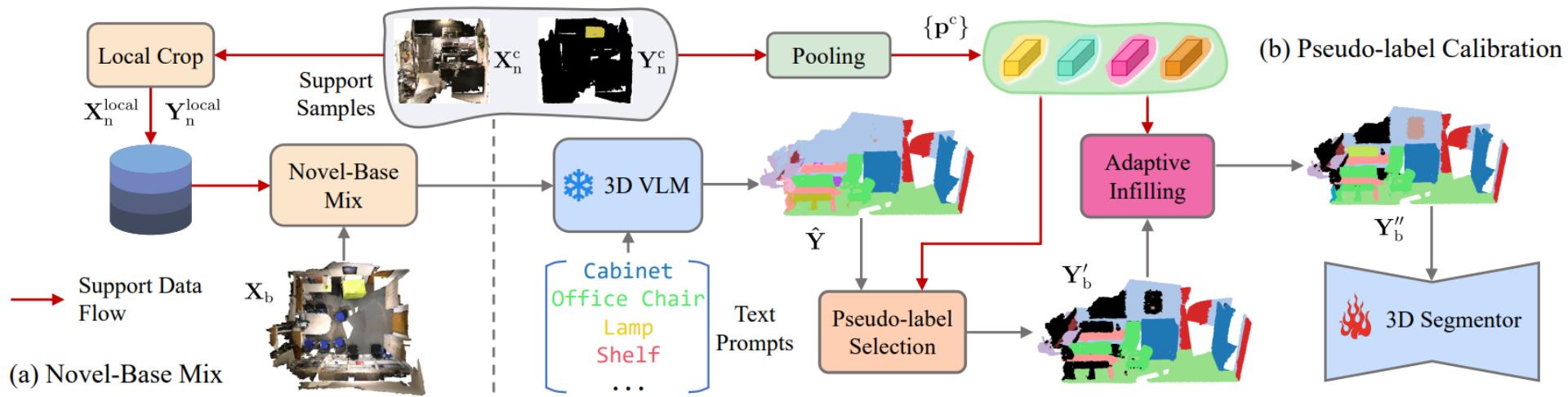
- Our framework addresses this limitation by leveraging the extensive open-world knowledge from 3D VLMs through pseudo-labels. We mitigate the noise inherent in 3D VLMs by calibrating their raw pseudo-labels with precise few-shot samples, thereby effectively expanding novel class knowledge while ensuring reliability.



- 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”. **IEEE CVPR, 2025**.

# GFS-PCS with 3D Vision-Language Models (3D VLMs)

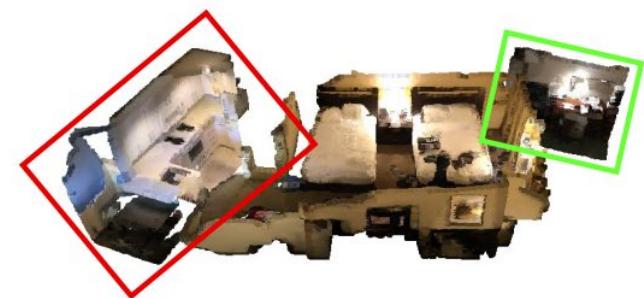
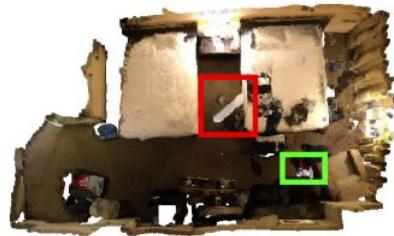
- Given an input point cloud  $\mathbf{X}_b$ , we apply a novel-base mix to embed support samples into the training scene while preserving essential context. The scene is then processed by a 3D VLM, using all class names as prompts to generate raw predictions  $\hat{\mathbf{Y}}$ . Leveraging support prototypes  $\{\mathbf{p}^c\}$ , the raw predictions undergo pseudo-label selection to filter out noisy regions, followed by adaptive infilling to label the filtered, unlabeled areas, yielding refined supervision  $\mathbf{Y}_b''$  for training the 3D segmentor.



- 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”. **IEEE CVPR, 2025**.

# GFS-PCS with 3D Vision-Language Models (3D VLMs)

- Visual illustration of mixing strategies. The red and green boxes represent the two novel samples mixed into the scene.



Instance Mix

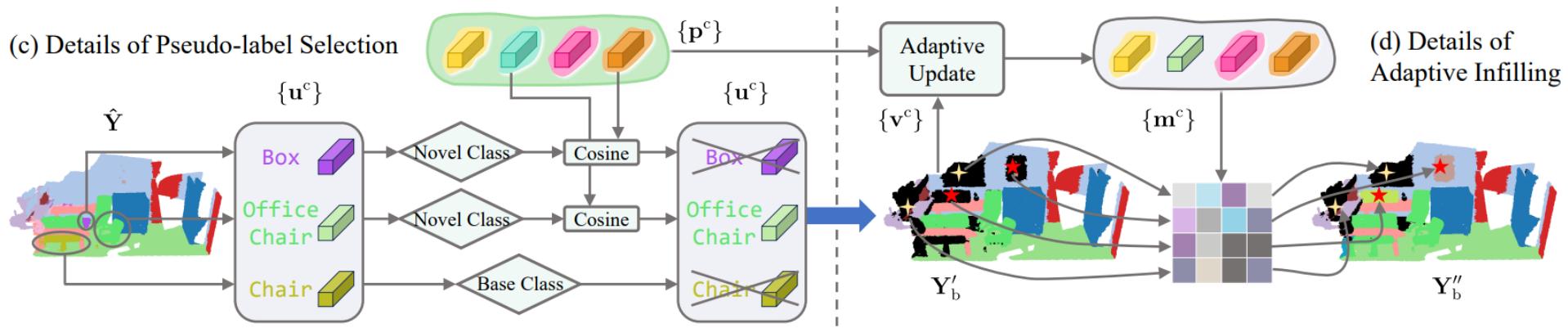
Mix3D

Novel-Base Mix (Ours)

- 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”. **IEEE CVPR, 2025**.

# GFS-PCS with 3D Vision-Language Models (3D VLMs)

- (c) (d) illustrate the details of the pseudo-label selection and adaptive infilling processes, respectively.



- We further propose two new and more challenging evaluation benchmarks based on ScanNet200 and ScanNet++ datasets.

| Dataset    | Base | Novel | Max (F)    | Min (F)    | Max (P)       | Min (P)    |
|------------|------|-------|------------|------------|---------------|------------|
| S3DIS      | 7    | 6     | 185        | 29         | 59,929        | 30,013     |
| ScanNet    | 13   | 6     | 411        | 133        | 4,479         | 1,148      |
| ScanNet200 | 12   | 45    | <b>733</b> | <b>102</b> | 12,641        | 279        |
| ScanNet++  | 12   | 18    | 143        | 82         | <b>84,375</b> | <b>604</b> |

- 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”. **IEEE CVPR, 2025**.

# Quantitative Comparison

| Method           | 5-shot       |              |              |              | 1-shot       |              |              |              |
|------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                  | mIoU-B       | mIoU-N       | mIoU-A       | HM           | mIoU-B       | mIoU-N       | mIoU-A       | HM           |
| Fully Supervised | 68.70        | 39.32        | 45.51        | 50.02        | 68.70        | 39.32        | 45.51        | 50.02        |
| PIFS [4]         | 28.78        | 3.82         | 9.07         | 6.71         | 17.84        | 2.87         | 6.02         | 4.88         |
| attMPTI [77]     | 37.13        | 4.99         | 11.76        | 8.79         | 54.84        | 3.28         | 14.14        | 6.17         |
| COSeg [2]        | 57.67        | 5.21         | 16.25        | 9.54         | 47.03        | 4.03         | 13.09        | 7.42         |
| GW [66]          | 59.28        | 8.30         | 19.03        | 14.55        | 55.23        | 6.47         | 16.74        | 11.56        |
| GFS-VL (ours)    | <b>67.57</b> | <b>31.67</b> | <b>39.23</b> | <b>43.12</b> | <b>68.48</b> | <b>29.18</b> | <b>37.45</b> | <b>40.92</b> |

Table 2. Comparisons of our method with baselines on the new ScanNet200 benchmark. The best results are highlighted in **bold**.

| Method           | 5-shot       |              |              |              | 1-shot       |              |              |              |
|------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                  | mIoU-B       | mIoU-N       | mIoU-A       | HM           | mIoU-B       | mIoU-N       | mIoU-A       | HM           |
| Fully Supervised | 65.45        | 37.24        | 48.53        | 47.47        | 65.45        | 37.24        | 48.53        | 47.47        |
| PIFS [4]         | 39.98        | 5.74         | 19.44        | 10.03        | 36.66        | 4.95         | 17.63        | 8.71         |
| attMPTI [77]     | 55.89        | 4.19         | 24.87        | 7.78         | 53.16        | 3.55         | 23.40        | 6.66         |
| COSeg [2]        | 59.34        | 6.96         | 27.91        | 12.45        | 58.49        | 6.24         | 27.14        | 11.26        |
| GW [66]          | 51.35        | 11.03        | 27.16        | 18.15        | 46.71        | 6.63         | 22.66        | 11.59        |
| GFS-VL (ours)    | <b>60.05</b> | <b>21.66</b> | <b>37.02</b> | <b>31.82</b> | <b>61.39</b> | <b>19.42</b> | <b>36.21</b> | <b>29.47</b> |

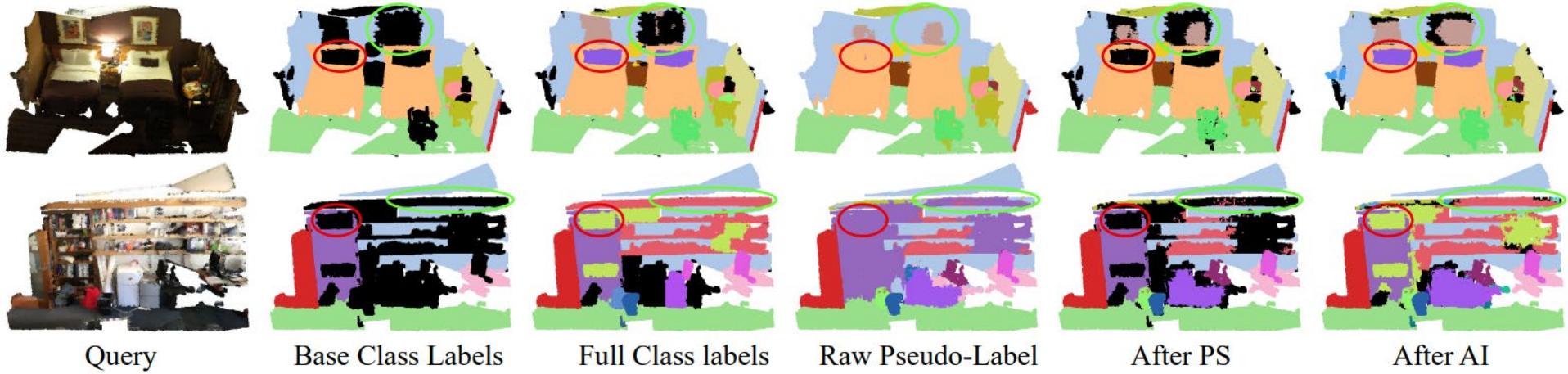
Table 3. Comparisons of our method with baselines on the new ScanNet++ benchmark. The best results are highlighted in **bold**.

| Method           | 5-shot       |              |              |              | 1-shot       |              |              |              |
|------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                  | mIoU-B       | mIoU-N       | mIoU-A       | HM           | mIoU-B       | mIoU-N       | mIoU-A       | HM           |
| Fully Supervised | 78.71        | 60.37        | 72.91        | 68.33        | 78.71        | 60.37        | 72.91        | 68.33        |
| attMPTI [77]     | 16.31        | 3.12         | 12.35        | 5.21         | 12.97        | 1.62         | 9.57         | 2.88         |
| PIFS [4]         | 35.14        | 3.21         | 25.56        | 5.88         | 35.80        | 2.54         | 25.82        | 4.75         |
| CAPL [56]        | 38.22        | 14.39        | 31.07        | 20.88        | 38.70        | 10.59        | 30.27        | 16.53        |
| GW [66]          | 40.18        | 18.58        | 33.70        | 25.39        | 40.06        | 14.78        | 32.47        | 21.55        |
| GFS-VL (ours)    | <b>78.30</b> | <b>51.22</b> | <b>69.75</b> | <b>61.91</b> | <b>78.56</b> | <b>49.72</b> | <b>69.45</b> | <b>60.88</b> |

Table 4. Comparisons of our method with baselines on the old ScanNet benchmark. The best results are highlighted in **bold**.

# Qualitative Comparison

- Visualization of the improvements in pseudo-label quality after applying Pseudo-label Selection (PS) and Adaptive Infilling (AI). Note that AI effectively discovers missed novel classes in the red circles and completes partial pseudo-labels in the green circles.



- 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”. **IEEE CVPR, 2025**.



SCAN ME

[https://yun-liu.github.io/materials/Slides\\_FS-PCS.pdf](https://yun-liu.github.io/materials/Slides_FS-PCS.pdf)

Thank you!