Supplementary Materials for Localization and Expansion: A Decoupled Framework for Point Cloud Few-shot Semantic Segmentation

Zhaoyang Li[1⋆], Yuan Wang[1⋆], Wangkai Li[1], Rui Sun[1], and Tianzhu Zhang [1,2†]

[1] MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
[2] Deep Space Exploration Laboratory
{lizhaoyang,wy2016,lwklwk,issunrui}@mail.ustc.edu.cn, tzzhang@ustc.edu.cn

In the supplementary material, we first introduce more details of the framework (Sec. 1). Subsequently, we provide additional method details (Sec. 2), then show the dataset partitioning details (Sec. 3), and conduct additional experiments (Sec. 4). Then, we present more visualization results (Sec. 5). Finally, we have included some additional discussions(Sec. 6)).

1 More Framework Details

1.1 Embedding Network

The architecture of the embedding network is kept the same as [7], which is composed of three modules: *feature extractor*, *attention learner* and *metric learner*. We also adopt DGCNN [6] as the backbone of our feature extractor. The input support and query point clouds are first processed through the embedding network, which outputs support features and query features for executing subsequent operations.

1.2 Transductive Inference

In the final prediction phase, we employed a transductive inference approach similar to that used in [7] to obtain the final predictions for the query. Specifically, this approach involves using transductive label propagation to construct a graph that includes both labeled multi-prototypes and unlabeled query points. Labels are then propagated within this graph utilizing a random walk method, thereby facilitating the determination of the final query predictions. The hyperparameter settings are consistent with those used in [7]. Contrary to [7] approach, in our method for generating multiple prototypes, the foreground prototypes are derived not only from the foreground points of the support set but also from the foreground target regions we excavate in the query images. Moreover, after generating the background prototypes, we further employ a mask cross-attention mechanism to adapt them to the background of the query images.

Table 1: Test class names for each split of S3DIS and ScanNet.

	split=0	split=1
S3DIS	beam, board, bookcase, ceiling, chair, column	door, floor, sofa, table, wall, window
ScanNet	otherfurniture, picture, refrigerator, show curtain, sink, sofa, table, toilet, wall, window	bathtub, bed, bookshelf, cabinet, chair, counter, curtain, desk, door, floor

2 Application of Mask Cross Attention (MCA) Mechanism to Background Prototypes

We obtain L background prototypes $\mathbf{P}_b \in \mathbb{R}^{L\times d}$ by employing K-means on support background features, as multiple prototypes enhance the representation of complex and cluttered backgrounds. But directly utilizing the background prototype $\mathbf{P}_b$ to guide the segmentation of query backgrounds can lead to suboptimal results due to contextual gaps between the support and query. Consequently, we use the mask cross attention (MCA) mechanism that leverages background features within the query to guide $\mathbf{P}_b$ in adapting to the query's background. In summary, for background prototype $\mathbf{P}_b$, query features $\mathbf{F}_q$, and the obtained target area $\mathbf{M}_{e\theta}$, we perform a mask cross attention operation, akin to Mask2Former [2], where *queries* ($\mathbf{Q}$) is derived from $\mathbf{P}_b$, *keys* ($\mathbf{K}$) and *values* ($\mathbf{V}$) are sourced from the query features $\mathbf{F}_q$, $\mathbf{M}_{e\theta}$ is utilized to execute a masking operation, formally:

$$\mathbf{Q} = \mathbf{P_b}\mathbf{W}^{\mathcal{Q}}, \quad \mathbf{K}_j = \mathbf{f}_j\mathbf{W}^{\mathcal{K}}, \quad \mathbf{V}_j = \mathbf{f}_j\mathbf{W}^{\mathcal{V}}, \tag{1}$$

among which, $\mathbf{W}^{\mathcal{Q}} \in \mathbb{R}^{d\times d_k}$, $\mathbf{W}^{\mathcal{K}} \in \mathbb{R}^{d\times d_k}$,$\mathbf{W}^{\mathcal{V}} \in \mathbb{R}^{d\times d_v}$ are linear projections. Our masked attention modulates the attention matrix via

$$\mathbf{P}_b = \mathrm{softmax}(\mathbf{M} + \mathbf{Q}\mathbf{K}^{\mathrm{T}})\mathbf{V}. \tag{2}$$

Moreover, the attention mask $\boldsymbol{\mathcal{M}}$ at coordinate positions i is

$$\boldsymbol{\mathcal{M}}(i) = \begin{cases} 0 & \text{if } \mathbf{M}_{e\theta}(i) = 0 \\ -\infty & \text{otherwise} \end{cases}. \tag{3}$$

Through the MCA operation, we confine the attention to the query background regions, enabling P_b to adapt more effectively to the background of the query while mitigating interference from foreground points.

3 Dataset Split

The class names included in each split of the S3DIS [1] and ScanNet [3] datasets are detailed in Table 1, presenting a clear categorization of the datasets.

4 Additional Ablations.

4.1 Effectiveness of the Mask Cross Attention Mechanism (MCA) to Background Prototypes.

We conduct experiments to compare the performance of using MCA to $\mathbf{P}_b$ against two alternatives: 1) the direct use of the original $\mathbf{P}_b$ (Experiment a); and 2) the implementation of cross-attention between $\mathbf{P}_b$ and the query feature $\mathbf{F}_q$, without using a masking mechanism (Experiment b) where *queries* ($\mathbf{Q}$) is derived from P_b, while the *keys* ($\mathbf{K}$) and *values* ($\mathbf{V}$) are sourced from the query features $\mathbf{F}_q$. As illustrated in Table 2, our MCA approach (Experiment c) surpasses these other methods, underscoring the effectiveness of our proposed strategy.

Table 2: Results of different methods for processing $\mathbf{P}_b$ under 1-way 1-shot setting on S3DIS (S^0).

Method	mIoU(%)
(a) Directly using $\mathbf{P}_b$	75.87
(b) Cross-attention between $\mathbf{P}_b$ and $\mathbf{F}_q$	75.80
(c) MCA	**76.54**

4.2 Threshold parameters.

The threshold hyperparameter is shared across datasets. Additional experiments on the threshold for the ScanNet dataset are shown in Fig. 2.

4.3 Qualitative results.

To better demonstrate the role and effectiveness of each module we proposed, we add qualitative results including an ablation study where each of the 2 modules is turned off, as shown in Fig. 1.

† Corresponding author
⋆ Equal contribution

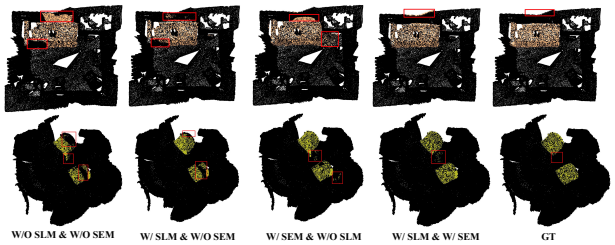

Fig. 1

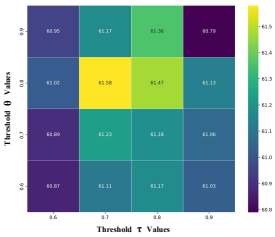

Fig. 2

4.4 Loss functions.

To explore the impact of the weight parameter, λ, within our loss equation, defined as $Loss = Loss_{CE} + \lambda Loss_{self}$, a series of experiments was conducted on the S3DIS dataset, adhering to the 1-way 1-shot S_0 setting. The findings, detailed in Table 3, indicate that a λ value set to 1 optimizes performance.

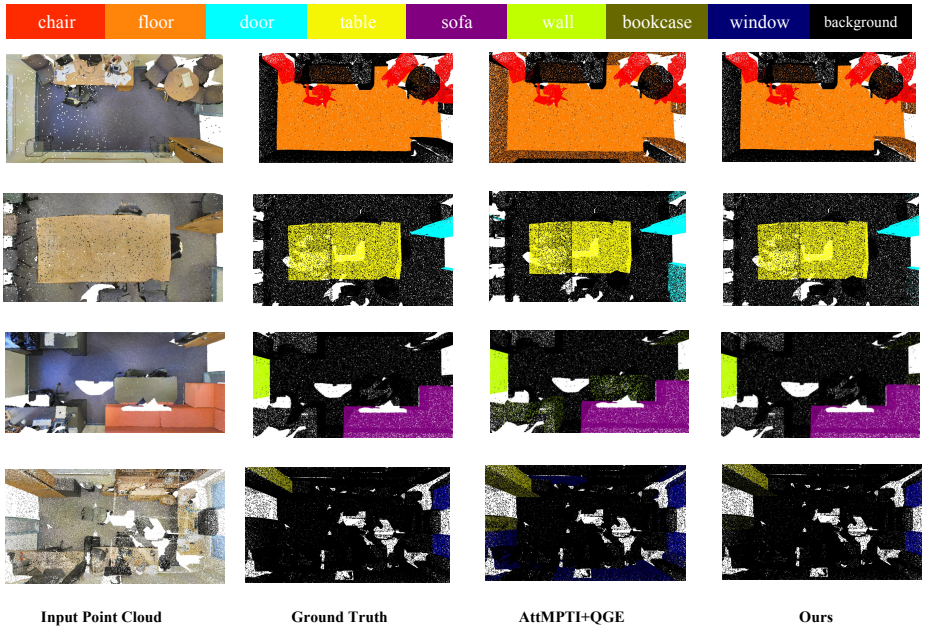

Fig. 3: qualitative results of our method in a 2-way 1-shot setting on the S3DIS [1] dataset, in comparison to the ground truth and the AttMPTI+QGE approach. Four combinations of 2-way are illustrated from the top to bottom rows, i.e., "*chair, floor*" (first row), "*door, table*" (second row), "*sofa, wall*" (third row), and "*bookcase, window*" (last row).

Table 3: Results of different loss functions under 1-way 1-shot setting on S3DIS (S^0).

Loss Function	mIoU(%)
$Loss = L_{CE}$	75.83
$Loss = L_{CE} + 0.2 * L_{self}$	75.97
$Loss = L_{CE} + L_{self}$	**76.54**
$Loss = L_{CE} + 2 * L_{self}$	76.29

5 Qualitative Results

From the visual results presented in Fig 3 and Fig 4, it is evident that our method demonstrates enhanced precision in distinguishing target category objects amidst multiple entities. It effectively minimizes interference from other similar objects, thereby yielding more accurate segmentation outcomes.

6 Discussions.

We follow QGE [5] to adopt 1&2 way settings as they can well represent single-way and multi-way scenarios (*i.e.*, simple and complex scenarios). To make a more comprehensive comparison, we add a 3-way experiment for S3DIS and ScanNet as shown in Tab. 5 and Tab. 6, which further demonstrate the superiority of our method in more complex scenarios. In Tab. 4, we quantitatively show that our model has a negligible FLOPS and inference time increment compared to the SOTA method (QGE).

Table 4

Methods	#Params	FLOPs	Avg. time
AttMPTI	357.82K	7.78G	0.1724
QGE	537.66K	8.14G	0.1818
Ours	542.98K	8.16G	0.2041

Table 5

Method	3-way					
	1-shot			5-shot		
	S^0	S^1	Mean	S^0	S^1	Mean
Ours	53.09	58.93	56.01	62.06	64.3	63.18

Table 6

Method	3-way					
	1-shot			5-shot		
	S^0	S^1	Mean	S^0	S^1	Mean
Ours	45.7	48.33	47.02	52.42	57.55	54.99

Additionally, to further demonstrate the novelty of our method, we compared it with previous 2D few-shot segmentation methods that also use multiple prototypes. We follow official implementations of ASGNet [4] to conduct comparative experiments, showing that directly applying 2D-FSS techniques to 3D-FSS is not feasible. Tab. 7 quantitatively shows that our method tailored for 3D-FSS surpasses ASGNet by a large margin.

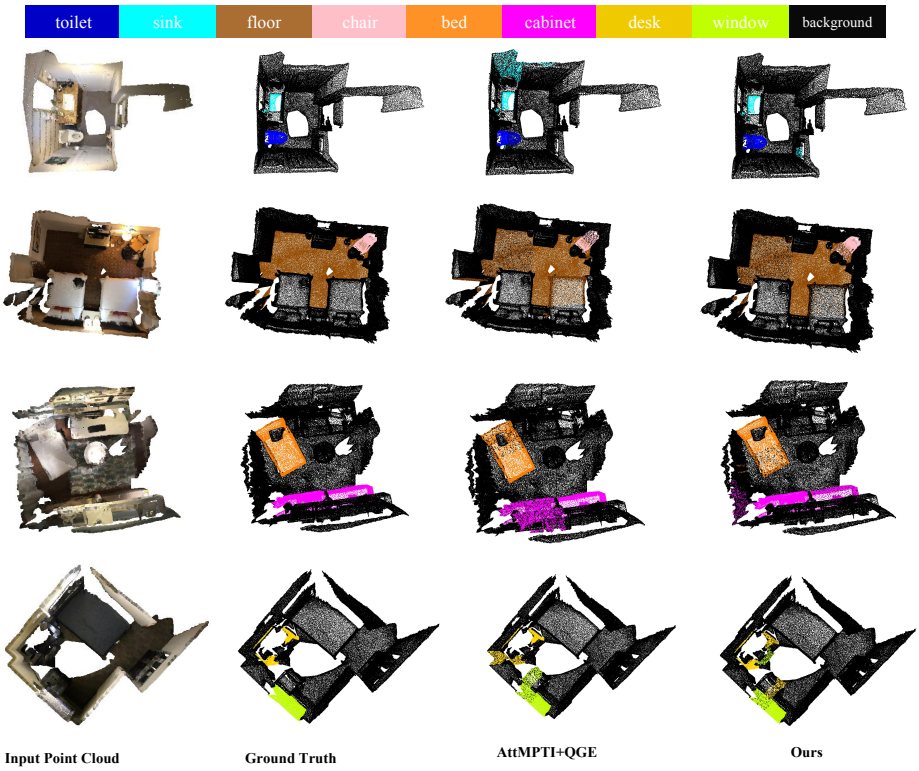

Fig. 4: qualitative results of our method in a 2-way 1-shot setting on the ScanNet [3] dataset, in comparison to the ground truth and the AttMPTI+QGE approach. Four combinations of 2-way are illustrated from the top to bottom rows, i.e., "*toilet, sink*" (first row), "*chair, floor*" (second row), "*bed, cabinet*" (third row), and "*desk, window*" (last row).

Table 7

Method	1-way						2-way					
	1-shot			5-shot			1-shot			5-shot		
	S^0	S^1	Mean	S^0	S^1	Mean	S^0	S^1	Mean	S^0	S^1	Mean
ASGNet	65.31	66.40	65.86	68.09	70.73	69.41	46.17	47.85	47.01	54.93	58.41	56.67
Ours	**76.54**	**78.8**	**77.67**	**83.15**	**83.23**	**83.19**	**61.34**	**63.58**	**62.46**	**67.92**	**74.49**	**71.36**

Limitations. The Self-Expansion Module (SEM) based on internal similarity may omit parts or include non-target areas when handling particularly heterogeneous objects.

References

1. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1534–1543 (2016)
2. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
3. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
4. Li, G., Jampani, V., Sevilla-Lara, L., Sun, D., Kim, J., Kim, J.: Adaptive prototype learning and allocation for few-shot segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8334–8343 (2021)
5. Ning, Z., Tian, Z., Lu, G., Pei, W.: Boosting few-shot 3d point cloud segmentation via query-guided enhancement. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1895–1904 (2023)
6. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics (tog) **38**(5), 1–12 (2019)
7. Zhao, N., Chua, T.S., Lee, G.H.: Few-shot 3d point cloud semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8873–8882 (2021)