Crossmodal Few-shot 3D Point Cloud Semantic Segmentation

Ziyu Zhao
University of South Carolina
ziyuz@email.sc.edu

Zhenyao Wu
University of South Carolina
zhenyao@email.sc.edu

Xinyi Wu
University of South Carolina
xinyiw@email.sc.edu

Canyu Zhang
University of South Carolina
canyu@email.sc.edu

Song Wang*
University of South Carolina
songwang@cec.sc.edu

ABSTRACT

Recently, few-shot 3D point cloud semantic segmentation methods have been introduced to mitigate the limitations of existing fully supervised approaches, *i.e.*, heavy dependence on labeled 3D data and poor capacity to generalize to new categories. However, those few-shot learning methods need one or few labeled data as support for testing. In practice, such data labeling usually requires manual annotation of large-scale points in 3D space, which can be very difficult and laborious. To address this problem, in this paper we introduce a novel crossmodal few-shot learning approach for 3D point cloud semantic segmentation. In this approach, the point cloud to be segmented is taken as query while one or few labeled 2D RGB images are taken as support to guide the segmentation of query. This way, we only need to annotate on a few 2D support images for the categories of interest. Specifically, we first convert the 2D support images into 3D point cloud format based on both appearance and the estimated depth information. We then introduce a co-embedding network for extracting the features of support and query, both from 3D point cloud format, to fill their domain gap. Finally, we compute the prototypes of support and employ cosine similarity between the prototypes and the query features for final segmentation. Experimental results on two widely-used benchmarks show that, with one or few labeled 2D images as support, our proposed method achieves competitive results against existing few-shot 3D point cloud semantic segmentation methods.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Crossmodal, 3D point cloud, Semantic segmentation

ACM Reference Format:
Ziyu Zhao, Zhenyao Wu, Xinyi Wu, Canyu Zhang, and Song Wang. 2022. Crossmodal Few-shot 3D Point Cloud Semantic Segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Oct. 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548251

*Corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal.

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548251

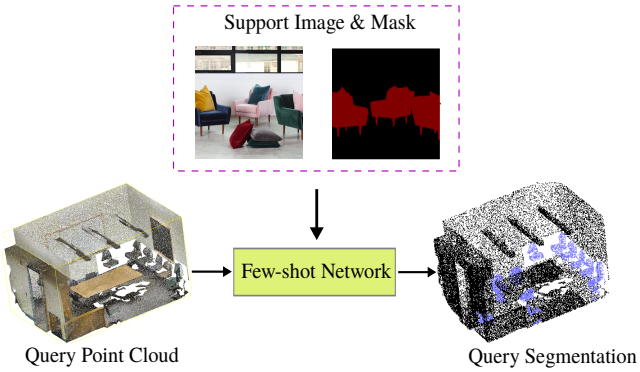

Figure 1: An illustration of our proposed crossmodal few-shot setting for 3D point cloud semantic segmentation. The labeled 2D image serves as the support and the query is a 3D point cloud. In this example, chair is the target category.

1 INTRODUCTION

3D point clouds recently have attracted a lot of attention of researchers because of its wide range of applications such as autonomous driving [4] , scene understanding [3] and indoor navigation [29]. 3D point cloud semantic segmentation is one of the fundamental tasks in computer vision by indicating the category of each point.

With the rapid development of deep learning techniques, existing methods [16, 19, 20, 26, 31, 33, 36, 40] have already achieved impressive results on many benchmarks. However, this success owes to large-scale labeled 3D point cloud datasets which usually are very expensive to obtain. Besides, the trained models may not extend well to novel categories.

These problems, recently, are mitigated by a few-shot learning-based method [40] proposed by Zhao *et al.* which only uses a few labeled 3D point clouds as exemplars to help segment the corresponding categories in the query point clouds. However, this method still requires a few labeled 3D point clouds for each category in both training and testing which serve as support exemplars. Given large-scale points in point clouds, manual labeling of just exemplars is still inconvenient and time-consuming. In the meantime 2D images are cheaper to obtain and easier to be annotated compared with 3D point clouds. This inspires us to use a few 2D labeled

images instead of labeled point clouds as the support. Based on this inspiration, in this paper, we propose a new crossmodal few-shot learning method for 3D point cloud semantic segmentation.

Specifically, we first collect a mini 2D image dataset to cover all the categories present in our query dataset (*i.e.*, the 3D point cloud dataset) by searching the internet. We then label these images for the categories of interest and take them as the support. The proposed few-shot learning pipeline is shown in Figure 1, where the category to be segmented in the query point cloud is labeled on the support image. To bridge the crossmodal gap between the 2D support and 3D query, we back-project the 2D image into 3D space by combining the RGB value and the estimated depth at every pixels. Our co-embedding network can further bridge the domain gap between the back-projected 3D data and the real query 3D point clouds. Experiments on S3DIS and ScanNet datasets show that our method can segment 3D point clouds for an unseen category by taking several labeled images as support and all the designed components in our network contribute to the performance improvement.

In this work, we make the following main contributions:

- We propose a novel 3D point cloud semantic segmentation method which uses labeled 2D images as support exemplars. To our knowledge, this is the first work to study the crossmodal few-shot learning between 2D images and 3D point clouds for 3D point cloud semantic segmentation.
- We construct a new 2D image support set to cover the categories to be segmented from the indoor scene 3D point cloud datasets. To bridge the gap between the two heterogeneous modalities, we convert the 2D images into 3D point cloud format by considering both the pixel intensity and the estimated depth of the images. We also introduce a co-embedding network to further fill the domain gap between the support and the query.
- We conduct comprehensive experiments to show that the proposed method can segment unseen categories of interest from 3D point clouds by using a few segmented 2D images as support.

2 RELATED WORK

2.1 3D Point Cloud Semantic Segmentation

Our work targets at the problem of 3D point cloud semantic segmentation. Most of existing solutions [13, 16, 19, 20, 26, 33, 36, 41] are fully-supervised by training on data with point-wise ground truths. PointNet [26] is a classic learning-based method for 3D semantic segmentation which learns a spatial encoding of each point and then aggregates all individual point features to a global point cloud signature. However, it fails to congregate the local feature from neighboring points. Wang *et al.* [33] introduce EdgeConv in their proposed Dynamic Graph Convolutional Neural Network (DGCNN) which not only incorporates local neighborhood information but learns global shape properties. Zhao *et al.* [40] incorporate this model into the 3D point cloud segmentation work as the backbone. Moreover, it combines the self-attention and metric learner module with the first EdgeConv layer to produce an integrated deep feature map embedded with the geometric and semantic information. Although this work helps alleviate the fully-supervised requirement and achieves reasonable evaluation performance, it is still based on training on the large amount of annotated 3D point clouds.

Different from the above methods, our method learns to segment 3D point clouds in the few-shot setting – only one or a few labeled exemplars are needed when segmenting for a unseen category.

2.2 Few-shot Semantic Segmentation

Another line of related works to is the few-shot semantic segmentation. Few-shot learning technique [18] is developed to help quickly generalize a trained model to novel categories using one or a few exemplar(s) and has been widely explored for image-level semantic segmentation. For example, several methods [8, 24, 32, 39] make use of the embedding network to match the common feature between support and query, where the support samples are usually aggregated into one global vector. Zhang *et al.* [38] propose to apply a masked average pooling to extract the foreground/background features from the support set. Most recently, few-shot learning technique has also been applied for 3D point cloud semantic segmentation by Zhao *et al.* [40]. Using a novel attention-aware multi-prototype transductive inference method , this few-shot learning method can significantly reduce the dependence on large-scale labeled 3D data. Nevertheless, this method still needs labeled 3D point clouds as support.

Different from [40], our method uses labeled 2D images as support to segment 3D point clouds.

2.3 Crossmodal Learning

Crossmodal learning occurs when the involved data come from different modalities. For instance, in [14, 15, 37], a natural-language description is used to instruct the localization and segmentation on an image. Ye *et al.* [35] also propose a natural language based crossmodal image segmentation approach with a self-attention module, which could efficiently control the long-range discrepancy and focus more on the regions of interest in the input image.

Complementary modality information has also been used to achieve 2D-3D crossmodal segmentation. A common way is to extract the 2D and 3D features with separate networks, and then filtering the dense 2D features into sparse point features to fill the modality gap. Peng *et al.* [25] enhance 2D-3D crossmodal segmentation by exploiting sparse-to-dense matching and learning strategy. Jaritz *et al.* [17] propose a crossmodal network with both unsupervised and semi-supervised domain adaptation, exploiting the modalities of RGB and LiDAR to make predictions on unlabeled 3D data. Different from these methods, we address point cloud semantic segmentation task under the few-shot learning setting, which can handle unseen categories during inference.

2.4 Single-image depth estimation

Single-image depth estimation [9, 12, 22, 23, 27, 27, 28] aims to regress depth by taking only a single image as input and it benefits many other tasks. Current state-of-the-art methods are mainly based on deep learning techniques. For instance, Chen *et al.* [4] improve object detection and semantic segmentation with a learnable depth sub-network. Bao *et al.* [2] explicitly detect the occlusion using depth information for video frame interpolation. Cheng *et al.* [6] propose a novel neural style transfer method with a depth-based

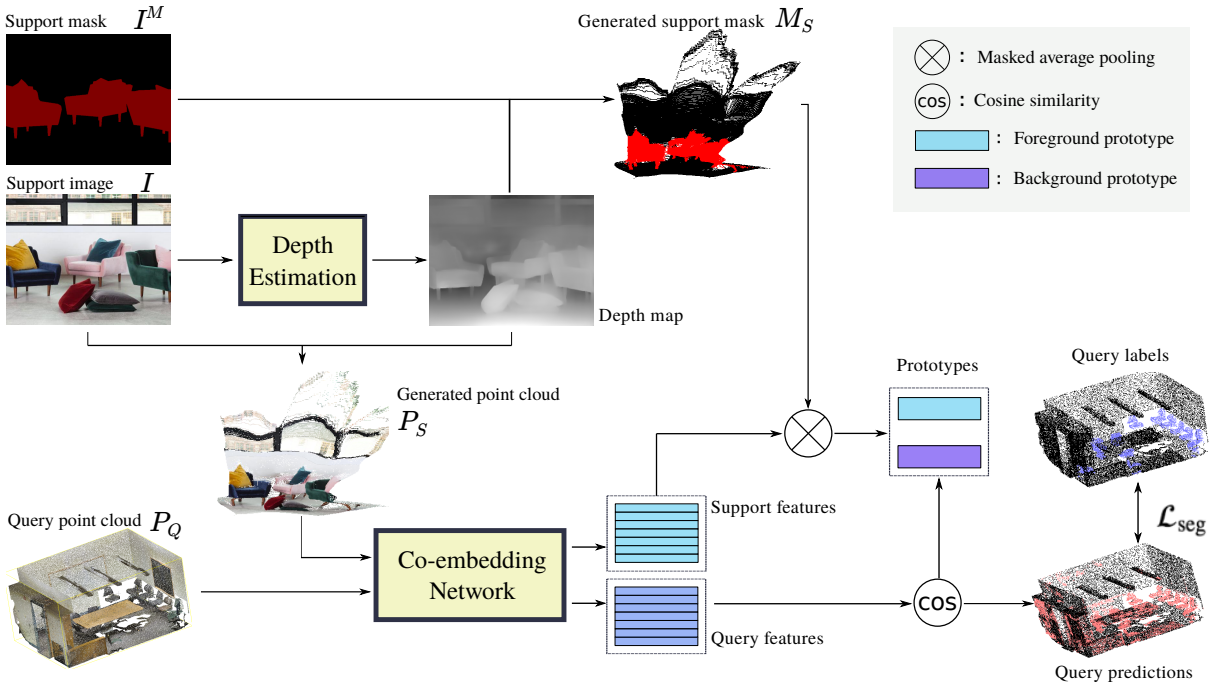

Figure 2: The pipeline of our proposed method under the 1-way 1-shot setting

global structure representation. He *et al.* [11] address the semantic segmentation and depth estimation jointly by exploiting the geometry constraint to simultaneously supervise these two tasks. In our work, we use a pre-trained depth estimator to infer the depth of our support images and convert the 2D images into 3D point cloud format for filling the crossmodal gap.

3 PROPOSED METHOD

3.1 Problem Definition

The proposed crossmodal few-shot point cloud semantic segmentation aims to categorize the 3D point clouds based on a few annotated 2D images as exemplars. To achieve this goal, we first collect a small set of 2D images from the internet based on all the categories that are involved in the 3D point cloud data and we annotate these corresponding categories on the 2D images. Each such 2D image and its annotated mask (I, I^M) are taken as a support sample and each 3D point cloud data P_Q serves as a query. For each query P_Q, its segmentation ground truth M_Q is used to provide supervision during training.

Both the support (2D) and query (3D) data are split into training set and testing set without overlapping categories, *i.e.*, the testing categories are not seen during training. For the K-way O-shot learning, the support set has O pair of (I, I^M) and there are in total K different categories used for testing. In the following section, we describe our proposed method mainly under 1-way 1-shot setting unless extra explanation.

3.2 Method Overview

Figure 2 illustrates the overall structure of our proposed method under the 1-way 1-shot setting. Given support image-mask pair (I, I^M) and query point cloud P_Q, we initially convert the 2D image and mask into pseudo 3D point cloud P_S for better matching the support and query features. This 2D-3D conversion is achieved by single-view depth estimation and geometric transformation, as detailed in Section 3.3. To extract features from pseudo point cloud P_S and query real point cloud P_Q, we propose a co-embedding network which is able to bridge their domain gap as well (Section 3.4). Finally, we compute the prototypes of the support data P_S and the segmentation results are obtained by calculating the cosine distance between the query feature vector and the prototypes, as detailed in Section 3.5.

3.3 2D to 3D Conversion

Given I and I^M as support data, the basic idea is to use 2D convolutions to extract deep features, while the query point cloud P_Q requires graph-based convolutions for processing. There will be a large modality gap between the image feature and the point cloud feature using different encoders. Therefore, we firstly consider aligning the modal of these two inputs. In this case, we consider transferring the 2D format data (I, I^M) to pseudo 3D point cloud (P_S, M_S) as support data before feature extraction. For each pixel in the 2D image I, we firstly estimate the depth value d via a pre-trained single-view depth estimator. Then, we back-project the pixel (u, v) into 3D space with specific camera's parameters, and

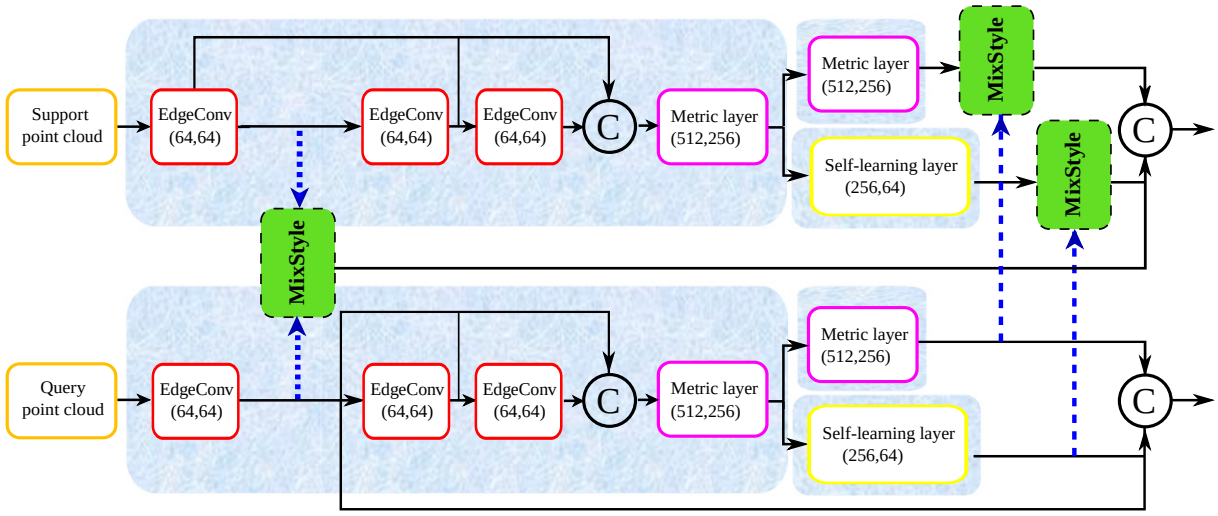

Figure 3: The structure of the proposed co-embedding network. It uses the query feature to modify the support feature via three "Mixstyle" modules, which are added after the first EdgeConv layer, the metric learner layer, and the self-learning layer, respectively.

the new coordinates (x, y, z) can be calculated by

$$z = \frac{d}{\sqrt{1 + z_x^2 + z_y^2}}, \qquad x = z_x z, \qquad y = z_y z,$$
$$z_x = \frac{(C_x - u)}{f}, \qquad z_y = \frac{(C_y - v)}{f}, \tag{1}$$

where (C_x, C_y) is the center of camera lens, and f denotes the focal length. In this paper, we choose the pre-trained model from [5] as single-view depth estimator and it can estimate accurate per-pixel depth map for pseudo point cloud generation. Note that we can use any other depth estimators if their results are reliable.

3.4 Co-Embedding Network

Let us first revisit the point cloud embedding network proposed in [40] for point cloud feature extraction, and it contains three parts. The first part consists of three EdgeConv layers [33] which are used for producing local geometric features and generating the semantic features. In order to further explore the interaction of semantic features among points, it employs a self-attention module [30] on the generated semantic features. Moreover, a metric learner module is used to enable the embedding network to get better adaptability for few-shot learning. Finally, the embedding network concatenates the outputs of the first EdgeConv layer, the self-learning module, and the metric learner module as the final output of the embedding network.

Given that our pseudo point cloud P_S is generated by image-based depth estimation, rather than the real-world scans, we propose a novel co-embedding network to fill the domain gap between the features of pseudo point cloud P_S and real query point cloud P_Q in feature extraction. Specifically, we integrate a "MixStyle" module [42] into the embedding network. As shown in Figure 3, one MixStyle module is added between the support branch and query branch after the first EdgeConv layer, the self-learning module and metric learner module, respectively. Let $f_s \in \mathbb{R}^{O\times C\times N}$ be one intermediate feature map of support branch, where O, C, N denote the number of shots, channel of features and number of input points. $\mu_{f_s} \in \mathbb{R}^C, \sigma_{f_s} \in \mathbb{R}^C$ denote the mean value and standard deviation within each channel of features. Specifically, for $c \in \{1, ..., C\}$,

$$\mu_{f_s^c} = \frac{1}{ON}\sum_{o=1}^{O}\sum_{n=1}^{N} f_s^{o,c,n} \tag{2}$$

and

$$\sigma_{f_s^c} = \sqrt{\frac{1}{ON}\sum_{o=1}^{O}\sum_{n=1}^{N}\left(f_s^{o,c,n} - \mu_{f_s^c}\right)^2}. \tag{3}$$

Similarly, we can obtain feature statistic $(\mu_{f_q}, \sigma_{f_q}) \in \mathbb{R}^C$ from query point cloud:

$$\mu_{f_q^c} = \frac{1}{ON}\sum_{o=1}^{O}\sum_{n=1}^{N} f_q^{o,c,n} \tag{4}$$

and

$$\sigma_{f_q^c} = \sqrt{\frac{1}{ON}\sum_{o=1}^{O}\sum_{n=1}^{N}\left(f_q^{o,c,n} - \mu_{f_q^c}\right)^2}. \tag{5}$$

Then, the MixStyle module combines their statistic attributes as:

$$\gamma_{mix} = \lambda\sigma_{f_s} + (1-\lambda)\sigma_{f_q}, \tag{6}$$

$$\beta_{mix} = \lambda\mu_{f_s} + (1-\lambda)\mu_{f_q}, \tag{7}$$

where λ is a random weight sampled from the uniform distribution, $\lambda \sim U(0, 1)$. Finally, we can obtain the aggregated support feature f_s' via:

$$f_s' = \gamma_{mix} \odot \frac{f_s - \mu_{f_s}}{\sigma_{f_s}} + \beta_{mix}. \tag{8}$$

3.5 Prototypes and Metric Learning

To build the correlation between the pseudo support point cloud and query point cloud, we adopt the prototype learner module [8] for our point cloud semantic segmentation task. The prototype of support feature is calculated by average pooling the feature in the same category. Specifically, given feature F_S from support branch, the foreground prototypes P_{fg} are computed via:

$$P_{fg}^{d_k,c} = \frac{1}{O}\sum_{o=1}^{O}\frac{\sum_N F_S^{o,c,n}\mathbb{1}[M_S^{o,n}=d_k]}{\sum_N \mathbb{1}[M_S^{o,n}=d_k]}, \tag{9}$$

where the d_i represents the target category and $\mathbb{1}[]$ is an indicator function, outputting 1 when the argument is true and 0 when false. The background prototypes are computed in similar fashion:

$$P_{bg}^{c} = \frac{1}{O}\sum_{o=1}^{O}\frac{\sum_N F_S^{o,c,n}\mathbb{1}[M_S^{o,n}\notin\mathbb{D}]}{\sum_N \mathbb{1}[M_S^{o,n}\notin\mathbb{D}]}, \tag{10}$$

where $\mathbb{D}=\{d_1,d_2,...,d_K\}$ is the set of all target categories.

Once the prototypes are obtained, we calculate the cosine distance between the query point cloud with each prototype. Finally, a softmax layer is applied over all distances to produce a probability map R for all categories plus the background, and the probability map for k-th category can be formulated as:

$$R_k = \frac{exp(-cossim(F_q,P_k))}{\sum_{P_k\in P} exp(-cossim(F_q,P_k))}, \tag{11}$$

where $P=\{P_{fg}\cup P_{bg}\}$ denotes all prototypes, F_q denotes the query feature, and $cossim$ is the function to measure the cosine similarity. During training, we utilize the cross entropy loss CE which measures the distance between the probability map R and query mask M_Q:

$$\mathcal{L}_{seg} = CE(R, M_Q). \tag{12}$$

4 EXPERIMENTS

4.1 Dataset

We evaluate the proposed model on the S3DIS [1] and ScanNet datasets [7]. The S3DIS dataset contains 272 point clouds of 6 large-scale indoor room scenes. Each point is annotated with one of the twelve semantic categories or the clutter category. The ScanNet dataset [7] contains 2.5 million indoor views in 1,513 scans, and each point is either annotated with one of the twenty semantic categories or unannotated as the background category. We examined massive existing 2D image datasets and none of them cover and well exhibit all the categories present in S3DIS or ScanNet. As mentioned earlier, we construct a new mini 2D image dataset as support for our work. Specifically, we search the internet and collect 13 sets of photos corresponding to 12 semantic categories in S3DIS[1] and 12 semantic categories in ScanNet[2]. Each set contains 5 different images and each image contains one or two semantic categories of interest. In addition, we annotate each pixel for all collected images according to the category of interest using binary masks.

[1] "$ceiling$", "$table$", "$wall$" ,"$floor$", "$chair$", "$door$", "$window$", "$sofa$", "$beam$" and "$column$".
[2] "$bookcase$", "$table$", "$wall$" ,"$floor$", "$chair$", "$door$", "$window$", "$sofa$", "$beam$" and "$column$".

4.2 Setup

In order to reduce the impact of our customized 2D image dataset on segmentation results, we randomly choose 6 categories of collected photos for training and the other 6 categories for testing as the support set. Accordingly, based on the training/testing categories, the point-cloud data in S3DIS and ScanNet are also evenly separated into 2 non-overlapped groups for training and testing. Each point cloud has its corresponding category specified (without considering the clutter category) when used as the query set. Considering a large number of points for each indoor scene, we pre-process the point clouds as in [26, 33]. Taking the S3DIS dataset for instance, in the query set, they divide all 272 point clouds of the indoor area into 7,547 blocks using a non-overlapping sliding window of 1m × 1m xy plane, and for each block, 2,048 points are randomly sampled such that there are at least 100 points for any target category. For the support set, we randomly sample 2,048 points for each support pseudo point cloud P_S, where 1,024 points are selected from the region of the target category and the other 1,024 points are sampled from the background. The corresponding mask M_S is processed in the same way.

4.3 Implementation Details

Training. Similar to [40], the weights in feature extractor module, self-attention and metric learner module are pre-trained on training set with batch size of 32, epochs of 100, and optimized by Adam with a learning rate of 0.0001. During the few-shot training, the model is optimized by Adam with a initial learning rate of 0.0001. All learning rates are decayed by 0.5 every 5000 iterations. The hyper-parameter f is set to 500.

Evaluation. We adopted the commonly used evaluation metric – mean Interaction over Union (mean-IoU). In our few-shot point cloud semantic segmentation task, mean-IoU is computed among two splits $split0$ and $split1$. The results of $split0$ is obtained on the model trained on the $split1$, and the results of $split1$ is obtained on the model trained on the $split0$. Essentially, the categories in two splits are randomly chosen and don't have any overlap. For the S3DIS dataset, $split0$ consists of "$ceiling$", "$table$", "$wall$" ,"$floor$" and "$chair$", and $split1$ consists of "$door$", "$window$", "$sofa$", "$beam$" and "$column$". For the ScanNet dataset, we replace the "$ceiling$" with "$bookcase$" category in $split0$ and stay the same in $split1$.

4.4 Baselines

Since we are the first work for crossmodal few-shot learning, we design two baselines for comparison with our method.

AffProto. In this baseline, we use an affine transformation method and regard all pixels in I as 3d points by adding a random z coordinate without depth estimation. Then the generated point cloud is sent into the embedding network for feature extraction, and the final prediction is produced by ProtoNet [40].

DepProto. This baseline is a degraded version of the proposed method, where the depth estimator module is kept, but the co-embedding network is removed – we use two share-weight embedding network [40] to extract the features of support and query point clouds, respectively.

Table 1: Quantitative results on S3DIS dataset.

Method	One-way						Two-way					
	One-shot			Five-shot			One-shot			Five-shot		
	*Split*0	*Split*1	*Mean*	*Split*0	*Split*1	*Mean*	*Split*0	*Split*1	*Mean*	*Split*0	*Split*1	*Mean*
AffProto	39.57	40.53	40.05	43.59	45.23	44.41	33.57	32.73	33.15	38.47	40.89	39.68
DepProto	45.89	44.10	44.50	53.23	55.89	54.06	40.27	38.58	39.43	44.44	43.28	43.86
Ours	**48.27**	**51.32**	**49.80**	**56.50**	**58.32**	**57.91**	**43.20**	**41.49**	**42.35**	**49.20**	**50.32**	**49.76**

Table 2: Quantitative results on ScanNet dataset.

Method	One-way						Two-way					
	One-shot			Five-shot			One-shot			Five-shot		
	*Split*0	*Split*1	*Mean*	*Split*0	*Split*1	*Mean*	*Split*0	*Split*1	*Mean*	*Split*0	*Split*1	*Mean*
AffProto	30.53	31.29	30.91	37.26	35.58	36.42	25.23	28.89	27.06	29.38	30.88	29.63
DepProto	38.56	39.10	38.83	42.89	41.29	41.59	33.52	31.58	32.05	38.26	39.25	38.76
Ours	**41.99**	**40.38**	**40.69**	**44.37**	**43.88**	**43.63**	**37.20**	**34.38**	**35.79**	**40.27**	**41.01**	**40.14**

4.5 Comparison with Baselines

Since we are the first to conduct the few-shot semantic segmentation on 3D point clouds through 2D image support, there are no prior experiment results that we could refer to. Therefore, we compare our method with these two baselines and the results are shown in Table 1 and Table 2 on the S3DIS and ScanNet dataset, respectively. We can see that, for all the methods, more shots and fewer categories lead to better segmentation performance – more shots provides more label information while fewer categories reduces the difficulty of training. Within these two baselines, **DepProto**, with the help of RGB-D information, outperforms **AffProto** in both 1-way and 2-way settings. Particularly, in both 1-way and 2-way settings, **DepProto** outperforms **AffProto** by around 15% and 10% respectively in S3DIS, and around 18% and 20% respectively in ScanNet. This indicates that, transferring 2D images to 3D point clouds by randomly setting a depth cannot provide an appropriate pseudo 3D point cloud for few-shot learning. The last row in both Table 1 and Table 2 show that our proposed method using the co-embedding network produces the best performance – mean-IoU improvement of 5% to 8% under different settings in both dataset. The superiority of our proposed method as compared with **DepProto** confirms the importance and contribution of feature fusion across the support and query

4.6 Qualitative Results

Qualitative results under 1-way 1-shot setting on the S3DIS dataset are shown in Figure 4. We compare the predictions of five unseen categories in several indoor scene point clouds using our proposed method and the **DepProto** baseline. As shown in Figure 4, in our method, the wall category (the third column of Figure 4) can be segmented clearly from other points with a relatively high accuracy by referring to the ground truth. Although the photo of chair category (the last column of Figure 4) we choose looks like sofa and shows large discrepancy with their corresponding category in the S3DIS dataset, our proposed method is still able to correctly recognize the basic shape from query while the prediction of **DepProto** contains strong noise – This demonstrates that the proposed co-embedding network has strong learning and generalization abilities in integrating features of support and query. Given limited support information in the 1-shot setting, the accuracy of query set segmentation is not perfect, but we still achieve visible segmentation improvement against **DepProto**, *e.g.* our method nicely segments the 'ceiling', 'table/desk', 'wall', 'floor', 'chair' in each scene in Figure 4.

We also provide the qualitative comparisons on ScanNet dataset in Figure 5. For the baseline methods, the results on ScanNet is not as good as the ones on S3DIS – the segmentation of several semantic categories is mixed with the background or other classes. For instance, it is hard to clearly separate the 'table/desk' segmentation from the 'chair' segmentation in the results produced by **DepProto** (the second column of Figure 5). However, our method could correctly segment most parts of these categories by fusing more semantic information of query features into the support. This verifies that the proposed method can well bridge the modality gap between 2D images and 3D point clouds.

4.7 Ablation Studies

4.7.1 ***Ablation Studies on the Co-embedding Network.*** To verify the design of the Co-embedding network, we conduct a comprehensive ablation study with several model variants as shown in Table 3. We insert the MixStyle module into different locations of network, i.e., just after the first EdgeConv layer (f_{Edge}), the self-attention module (f_{Att}) and/or the metric learner module (f_{MLP}). Besides, we use the MixStyle to modulate the features of different

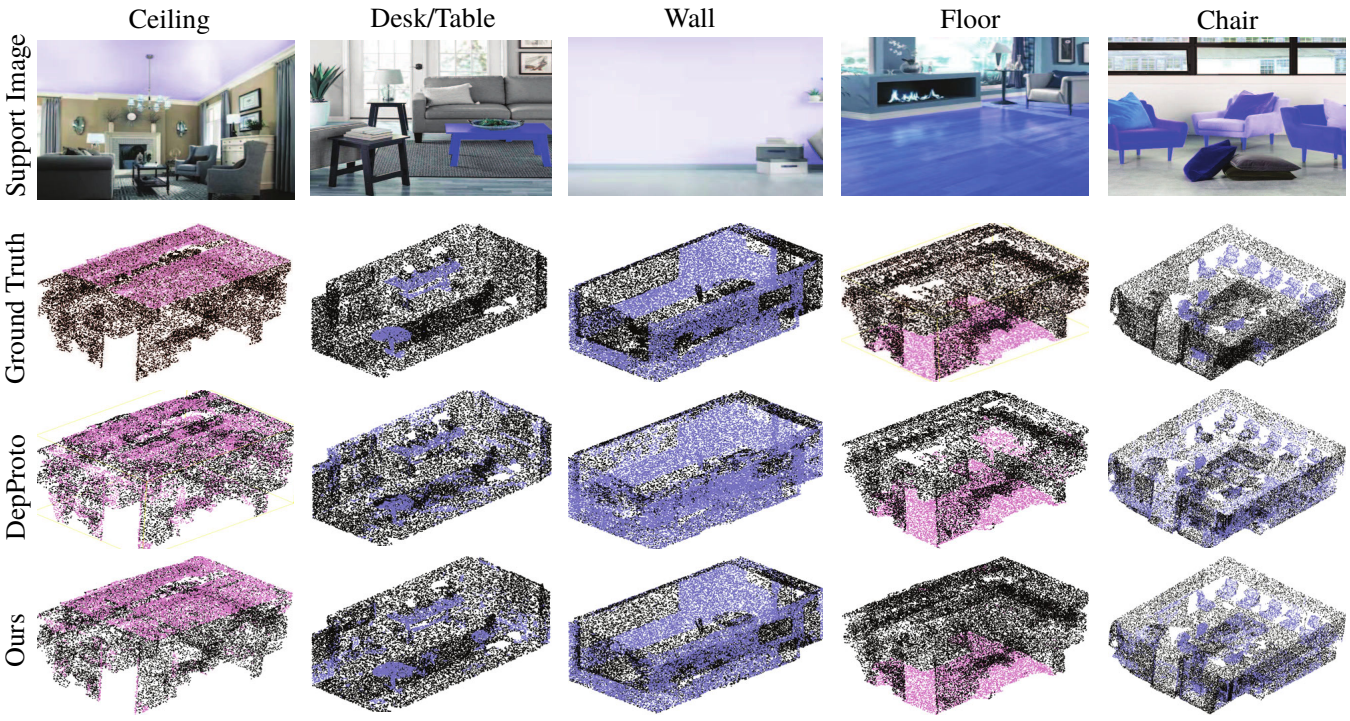

Figure 4: Qualitative results of our proposed method and DepProto for the crossmodal 1-way 1-shot point cloud semantic segmentation on the S3DIS dataset.

branches (support S and query Q). From Table 3, we can generally observe that using "MixStyle" for all three locations outperforms using it at only one location in terms of the mean-IoU. Moreover, we also find that using "MixStyle" to modulate the feature map of pseudo support yields better performance gain than query (up to around 8.3%).

Table 3: Results of inserting "MixStyle" at different locations under 1-way 5-shot setting on the S3DIS dataset.

f_{Edge}	f_{Att}	f_{MLP}	S	Q	*Split*0	*Split*1	*Mean*
✓			✓		47.27	46.23	46.75
	✓		✓		48.23	49.24	48.74
		✓	✓		50.81	51.26	50.04
✓				✓	48.63	47.14	47.89
	✓			✓	51.52	52.23	51.88
		✓		✓	52.89	53.14	53.02
✓	✓	✓	✓		**56.50**	**58.32**	**57.91**
✓	✓	✓		✓	48.56	47.23	47.90

4.7.2 ***Ablation Studies on the Depth Estimation Network*** . We study the effect of the single-image depth estimation network that is used for 2D to 3D Conversion in our proposed method. The studies are summarized in Table 4, where we compare the depth estimators of Megadepth [21], MonoDepth2 [10] and ManyDepth [34] in the inference stage. Figure 6 shows the generated pseudo support point clouds produced by these three estimators. Compared with DepProto, our method can increase the mean-IoU by 7.1%, 3.2% and 6.4% by using Megadepth, MonoDepth2 and ManyDepth, respectively. We also notice that our proposed method can achieve comparable results when using different depth estimators in inference. It shows the strong generalization ability of the proposed method.

Table 4: Ablation studies on different depth estimation networks under 1-way 5-shot setting on the S3DIS dataset.

Method		*Split*0	*Split*1	*Mean*
MegaDepth [21]	DepProto	53.23	55.89	54.06
	Ours	**56.50**	**58.32**	**57.91**
MonoDepth2 [10]	DepProto	50.27	50.19	50.73
	Ours	**53.50**	**51.21**	**52.36**
ManyDepth [34]	DepProto	48.26	49.08	48.67
	Ours	**52.37**	**51.25**	**51.81**

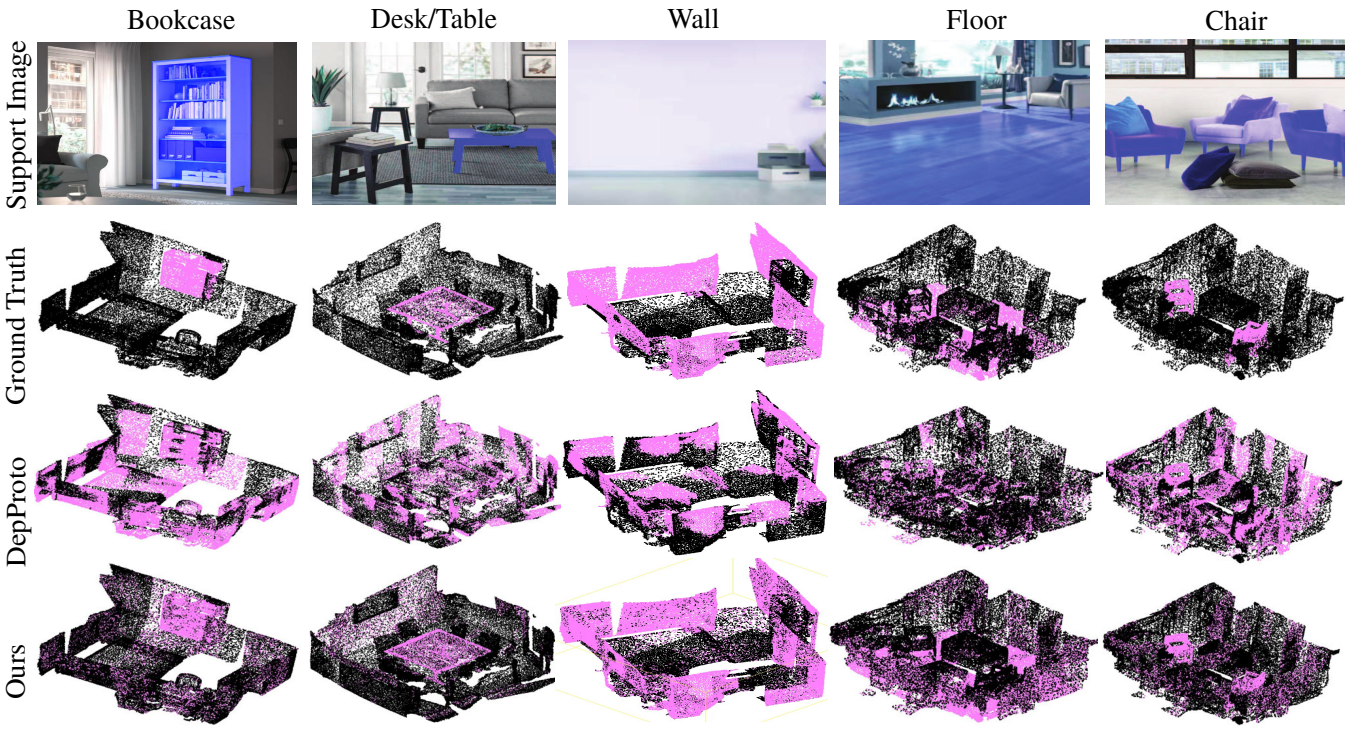

Figure 5: Qualitative results of our proposed method and DepProto for the crossmodal 1-way 1-shot point cloud semantic segmentation on the ScanNet dataset.

(a) MegaDepth (b) MonoDepth2 (c) ManyDepth

Figure 6: Qualitative results of generated support point cloud from different depth estimation.

4.7.3 ***Ablation Studies on the Hyper-Parameters*** . We also study the effect of focal length parameter f, and the results are shown in Table 5. The focal length f is used for 2D to 3D conversion (Eq.(1)). We evaluate our proposed method under 1-way 5-shot. From the Table 5 we can roughly conclude that when focal length approximates to 500, the mean-IoU can achieve the best.

Table 5: Results of studying parameter f under 1-way 5-shot setting on the S3DIS dataset.

f	200	400	500	600	800
S3DIS	55.28	56.21	**56.50**	54.19	54.79

5 LIMITATIONS

A number of limitations are still exposed over our method. Firstly, due to the complexity and diversity of point clouds, our customized training data could not perfectly adapt to all kinds of 3D data *e.g.* outdoor scenes scanned by different sensors. Secondly, since the monocular depth estimation may not perfectly recover the real 3D structure from a single image, it may mislead the following point cloud semantic segmentation during training and testing. Moreover, the same category in 2D images and 3D point clouds may still exhibit significant feature discrepancy, *e.g.* the appearances of "table" and "chair" shown in the Figure 1, which may not be fully addressable by the proposed method.

6 CONCLUSION

In this paper, we presented a novel crossmodal few-shot method for 3D point cloud semantic segmentation. Considering the high cost of 3D point-wise labeling, we used the labeled 2D RGB images as support to guide the semantic segmentation of 3D query point clouds. Specifically, we used the single image depth estimation to back-project the support from 2D to 3D, and introduced a co-embedding network to generalize deep features between support and query. Extensive visual comparisons and quantitative evaluation demonstrated that our proposed method performs well on the S3DIS and ScanNet datasets.

REFERENCES

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 2016. 3d semantic parsing of large-scale indoor spaces. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1534–1543.

[2] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. 2019. Depth-aware video frame interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3703–3712.

[3] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *IEEE International Conference on Computer Vision*. 9297–9307.

[4] Liangfu Chen, Zeng Yang, Jianjun Ma, and Zheng Luo. 2018. Driving scene perception network: Real-time joint detection, depth estimation and semantic segmentation. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 1283–1291.

[5] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. 2016. Single-image depth perception in the wild. *Advances in Neural Information Processing Systems* 29 (2016).

[6] Ming-Ming Cheng, Xiao-Chang Liu, Jie Wang, Shao-Ping Lu, Yu-Kun Lai, and Paul L Rosin. 2019. Structure-preserving neural style transfer. *IEEE Transactions on Image Processing* 29 (2019), 909–920.

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5828–5839.

[8] Nanqing Dong and Eric P Xing. 2018. Few-shot semantic segmentation with prototype learning.. In *BMVC*, Vol. 3.

[9] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems* 27 (2014).

[10] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. 2019. Digging into self-supervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3828–3838.

[11] Lei He, Jiwen Lu, Guanghui Wang, Shiyu Song, and Jie Zhou. 2021. SOSD-Net: Joint semantic object segmentation and depth estimation from monocular images. *Neurocomputing* 440 (2021), 251–263.

[12] Derek Hoiem, Alexei A Efros, and Martial Hebert. 2005. Geometric context from a single image. In *IEEE International Conference on Computer Vision*, Vol. 1. IEEE, 654–661.

[13] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. 2020. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition*. 11108–11117.

[14] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling relationships in referential expressions with compositional modular networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1115–1124.

[15] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4555–4564.

[16] Qiangui Huang, Weiyue Wang, and Ulrich Neumann. 2018. Recurrent slice networks for 3d segmentation of point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2626–2635.

[17] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Émilie Wirbel, and Patrick Pérez. 2021. Cross-modal learning for domain adaptation in 3d semantic segmentation. *arXiv preprint arXiv:2101.07253* (2021).

[18] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, Vol. 2. Lille, 0.

[19] Loic Landrieu and Martin Simonovsky. 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4558–4567.

[20] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. 2018. Pointcnn: Convolution on x-transformed points. *Advances in Neural Information Processing Systems* 31 (2018).

[21] Zhengqi Li and Noah Snavely. 2018. Megadepth: Learning single-view depth prediction from internet photos. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2041–2050.

[22] Fayao Liu, Chunhua Shen, and Guosheng Lin. 2015. Deep convolutional neural fields for depth estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5162–5170.

[23] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. 2015. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence* 38, 10 (2015), 2024–2039.

[24] Khoi Nguyen and Sinisa Todorovic. 2019. Feature weighting and boosting for few-shot segmentation. In *IEEE International Conference on Computer Vision*. 622–631.

[25] Duo Peng, Yinjie Lei, Wen Li, Pingping Zhang, and Yulan Guo. 2021. Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7108–7117.

[26] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 652–660.

[27] Anirban Roy and Sinisa Todorovic. 2016. Monocular depth estimation using neural regression forest. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5506–5514.

[28] Ashutosh Saxena, Sung Chung, and Andrew Ng. 2005. Learning depth from single monocular images. *Advances in Neural Information Processing Systems* 18 (2005).

[29] Daniel Teso-Fz-Betoño, Ekaitz Zulueta, Ander Sánchez-Chica, Unai Fernandez-Gamiz, and Aitor Saenz-Aguirre. 2020. Semantic segmentation to develop an indoor navigation system for an autonomous mobile robot. *Mathematics* 8, 5 (2020), 855.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).

[31] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in Neural Information Processing Systems* 29 (2016).

[32] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. 2019. Panet: Few-shot image semantic segmentation with prototype alignment. In *IEEE International Conference on Computer Vision*. 9197–9206.

[33] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. 2019. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)* 38, 5 (2019), 1–12.

[34] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. 2021. The temporal opportunist: Self-supervised multi-frame monocular depth. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1164–1174.

[35] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10502–10511.

[36] Xiaoqing Ye, Jiamao Li, Hexiao Huang, Liang Du, and Xiaolin Zhang. 2018. 3d recurrent neural networks with context fusion for point cloud semantic segmentation. In *European Conference on Computer Vision*. 403–417.

[37] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1307–1315.

[38] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. 2019. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *IEEE International Conference on Computer Vision*. 9587–9595.

[39] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. 2019. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5217–5226.

[40] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. 2021. Few-shot 3d point cloud semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 8873–8882.

[41] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. 2021. Ps^2-net: A locally and globally aware network for point-based semantic segmentation. In *International Conference on Pattern Recognition*. IEEE, 723–730.

[42] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. 2021. Mixstyle neural networks for domain generalization and adaptation. In *International Conference on Learning Representations*.