SPDC: A SUPER-POINT AND POINT COMBINING BASED DUAL-SCALE CONTRASTIVE LEARNING NETWORK FOR POINT CLOUD SEMANTIC SEGMENTATION

Shuai Zhang[1], Weihong Huang[1], Yiping Chen[1]*, Shuhang Zhang[1]*, Wuming Zhang[1], Jonathan Li[2]

[1] School of Geospatial Engineering and Science, Sun Yat-sen University, 519082 Zhuhai, China-
(zhangsh255, huangwh68)@mail2.sysu.edu.cn
[1] School of Geospatial Engineering and Science, Sun Yat-sen University, 519082 Zhuhai, China-
(chenyp79, zhangsh52, zhangwm25)@mail.sysu.edu.cn
[2] Geography and Environmental Management, University of Waterloo, N2L 3G1 ON, Canada-
junli@uwaterloo.ca

KEY WORDS: Point cloud semantic segmentation, Dual scale contrastive learning, Super point generation, Dynamic data augmentation.

ABSTRACT:

Semantic segmentation of point clouds is one of the fundamental tasks of point cloud processing and is the basis for other downstream tasks. Deep learning has become the main method to solve point cloud processing. Most existing 3D deep learning models require large amounts of point cloud data to drive them, but annotating the data requires significant time and economic costs. To address the problem of semantic segmentation requiring large amounts of annotated data for training, this paper proposes a **S**uper-point-level and **P**oint-level **D**ual-scale **C**ontrast learning network (**SPDC**). To solve the problem that contrastive learning is difficult to train and feature extraction is not sufficient, we introduce super-point maps to assist the network in feature extraction. We use a pre-trained super-point generation network to convert the original point cloud into a super-point map. A dynamic data augmentation(DDA) module is designed for the super-point maps for super-point-level contrastive learning. We map the extracted super-point-level features back to the original point-level scale and conduct secondary contrastive learning with the original point features. The whole feature extraction network is parameter sharing and to reduce the number of parameters we used the light-weight network DGCNN (encoder)+Self-attention as the backbone network. And we did a few-shot pre-training of the backbone network to make the network converge easily. Analogous to CutMix, we designed a new method for point cloud data augmentation called PointObjectMix (POM). This method solves the sample imbalance problem while preserving the overall characteristics of the objects in the scene. We conducted experiments on the S3DIS dataset and obtained 63.3% mIoU. We have also done a large number of ablation experiments to verify the effectiveness of the modules in our method. Experimental results show that our method outperforms the best-unsupervised network available.

1. INTRODUCTION

With the development of technology, two-dimensional computer vision data processing has gradually become unsatisfactory for real-world applications. As a fine-grained representation of 3D data, the point cloud has received increasing attention from researchers and several tasks about 3D perception have been widely studied, among which point cloud semantic segmentation has been a hot topic. Semantic segmentation of 3D point clouds aims to classify each individual point into a semantic label. While deep learning techniques have been applied in point cloud data, deep networks have become the main solution for point cloud semantic segmentation. Deep networks have been widely applied for this task and have achieved fine performance. Major approaches can be divided into point-based, voxel-based, and projection-based networks.

In the past few years, 3D feature and representation learning based on deep networks have made great progress. However, supervised 3D deep learning models require large amounts of annotated point cloud data to drive them, but annotating the data requires significant time and economic costs. As the self-supervised approach has been shown effective in 2D domains, 3D self-supervised methods that strip annotated data has gained increasing attention. For instance, it is an effective approach to build a contrastive learning framework to learn point features while only using the original data itself.

Another approach to process large-scale point data is to perform over-segmentation, which segments the points into super points with a less total number. During this process, points with similar geometric and semantic characteristics are divided into a cluster named super point, which is called the over-segmentation of point clouds. Traditional over-segmentation algorithms can be divided into cluster-based and graph-based methods. Most of the cluster-based methods are based on the ideas of K-Means. Graph-based methods consider each point as a node and construct edges using similarity and connectivity between points. These methods all try to over-segment point clouds according to certain criteria but rely on manual initialization and features. And performing over-segmentation using deep networks also starts attracting interest. The current major idea of using deep learning techniques to help over-segment points is to combine deep features with clustering or graph-cutting ideas, which is divided into two steps: first, we learn deep features through feature learning modules, and then we use clustering or graph cut methods to obtain the final super points.

In this paper, we propose a **S**uper-point-level and **P**oint-level

* : Corresponding author

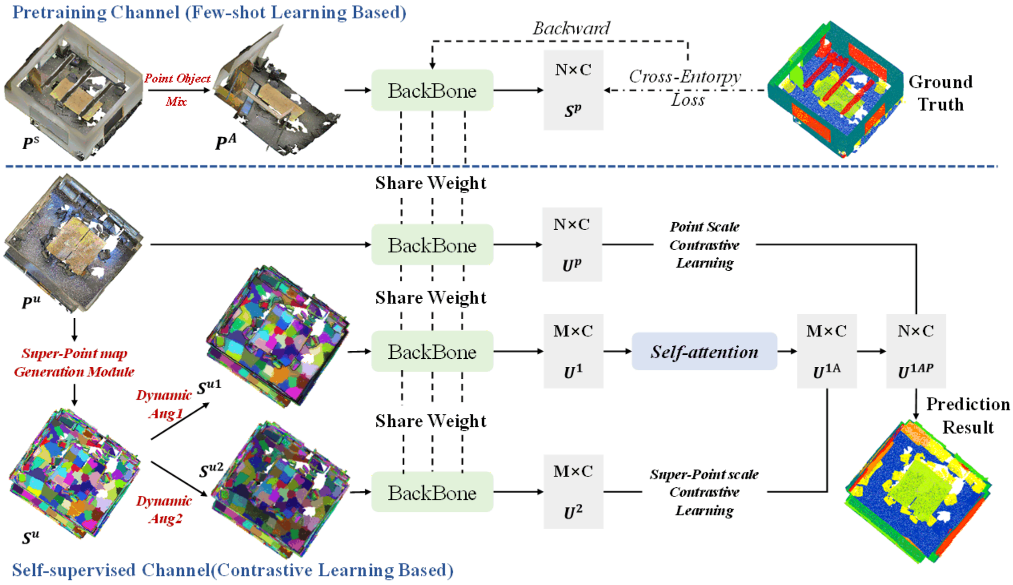

Figure 1. The Overall Architecture of The Proposed Method (SPDC)

Dual-scale **C**ontrast learning network (**SPDC**) to solve point cloud segmentation problem. Our main contribution can be summarized as follows:

1) We proposed a dual-scale contrastive learning network called SPDC, based on both a pretraining channel and a self-supervised model. The model trained by a fully supervised approach is transferred to unannotated data. Weight sharing is used to extract unannotated data features. Annotation-free generation of super point maps for multi-scale feature extraction and dual-scale comparative learning to correct the inhibitory nature of network transfer.

2) In the pre-training channel, to alleviate the gap between samples of different categories, we designed PointObject-Mix model for data augmentation in analogy to CutMix and PointCutMix.

3) In the self-supervised channel, we used a lightweight network model to generate super-point clusters and designed a dynamic data augmentation module for the super-point map to facilitate contrastive learning among super-point map features.

4) Our proposed method was experimented on the S3DIS dataset extensively and obtained the equivalent experimental accuracy as the fully supervised method. The SPDC optioned the 63.3% mIoU on S3DIS dataset, which is a state-of-art performance in the self-supervised point cloud semantic segmentation task.

2. RELATED WORKED

2.1 Fully Supervised 3D Semantic Segmentation Networks

Inspired by PointNet (Qi et al., 2017a) and PointNet++ (Qi et al., 2017b), MLP and max pooling layer can be directly used on irregular point data. RandLA-Net (Hu et al., 2020) utilizes random point sampling to efficiently learn features of large-scale datasets. SCF-Net (Fan et al., 2021) proposed Dual-Distance Attentive Pooling to learn spatial contextual features. Some other methods rely on the voxel data structure. VV-Net (Meng et al., 2019) takes each voxel grid as a unit and proposes a kernel-based interpolated variational autoencoder framework to extract local information. There are also methods combining point and voxel structures. For instance, point-voxel CNN framework (Liu et al., 2019) predicts the affinity of each voxel grid. Projection-based methods project point cloud data into 2D multi-view or spherical images and then employ well-established 2D CNN structures. MVCNN (Su et al., 2015) and RangeNet++ (Milioto et al., 2019) are two representative works.

2.2 Point Cloud Oversegmentation

VCCS (Papon et al., 2013) constructs voxel data structure for the point cloud and performs super-voxel division based on the adjacency of voxels. And it is the pioneering over-segmentation method based on clustering. VCCS-knn method (Sha et al., 2020) improves the neighboring searching methods on the basis of VCCS, which better ensured that the super points obtained by segmentation would not destroy the boundaries between real objects. PCLV method (Ben-Shabat et al., 2018) extends the graph cut problem in 2D images to point cloud data, and realizes the over-segmentation of point clouds. In the SPG network (Landrieu and Simonovsky, 2018), manually extracted point features are used, and the nearest neighbors of points are used to construct edges. The problem is turned into the minimum cut problem of the graph.

Two representative deep learning-based works are SSP (Landrieu and Boussaha, 2019) and SPNet (Hui et al., 2021). SSP network implements an end-to-end graph-based super voxel segmentation method. SPNet implements a differentiable version of SLIC for super voxel segmentation. These two networks are able to generate super points with self-adaptive numbers and size and have better edge-preserving properties.

2.3 Self-supervised Networks on Point Cloud

Self-supervised methods provide a new way to avoid the larger amount of annotated data and can improve the efficiency of

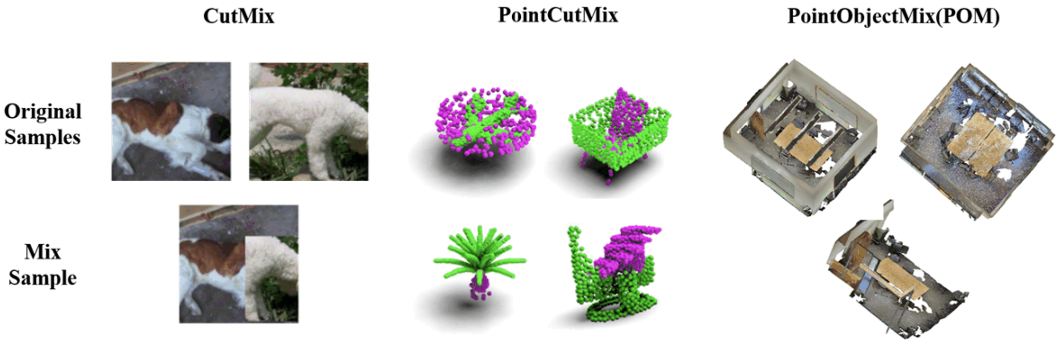

Figure 2. PointObjectMix Data Augmentation

models. Some works apply generative approaches to learn high-level representation from 3D point clouds. For instance, OcCo (Wang et al., 2021) selects completion as a pre-text task. Point-Bert (Yu et al., 2022) also learns by completing of masked area, while using a Transformer structure. Other work learns context information rather than trying to generate complete data. PointContrast (Xie et al., 2020) proposes a contrastive learning framework to learn representation from two views of the same scene. Spatio-temporal Representation Learning Network (Huang et al., 2021) learns spatial and temporal structures from two neighboring point cloud frames while trying to minimize the MSE between the learned features of the pair. Self-Correction (Chen et al., 2021) is a hybrid method that learns shape features by distinguishing and restoring destroyed objects. Motivated by PointContrast, we employ a contrastive learning framework for unsupervised representation learning.

3. PROPOSED METHOD

In this work, we introduce super-point-level point cloud over-segmentation and then construct a dual-scale contrastive learning network based on a mixture of points and super-points. We also designed a pre-training channel with few-shot learning to provide the network with initial values for a specific semantic segmentation downstream task. In order to adapt to the contrastive learning network, we designed two types of data augmentation modules corresponding to the characteristics of point and super-point maps, which were the PointObjectMix (POM) module and the Dynamic Data Augmentation(DDA) module. In this section, we will introduce our network in two main parts: pre-trained channels(Sec. 3.1) and self-supervised channels(Sec. 3.2). The overview of the proposed method is shown in Figure 1.

3.1 Pretraining Channel

3.1.1 Point Object Mix: In the pre-training channel, similar to other networks, we use Ground Truth to train the backbone network. But initially, in order to increase the sample size and balance the number of samples in different categories, we usually use some data augmentation methods, such as random panning, rotation, etc. However, for point cloud data with rotational invariance, the traditional rigid transformation to data augmentation will not work well. For advanced means of data enhancement, mixed sample data augmentation(MSDA) has received more attention in 2D image processing. Among the most widely utilized methods are MixUp(Zhang et al., 2017) and CutMix(Yun et al., 2019). MixUp interpolates between sample pixels to create more training samples. And Cutmix is used to create more training data by inserting parts of other scenes into the sample to be processed. Meanwhile, based on these two ideas, PointMixUp(Chen et al., 2020b) and PointCutMix(Zhang et al., 2022) also appear in point clouds. However, for point cloud data with semantic information, random cutting and stitching will destroy the inherent structural semantic information of the point cloud. Therefore, we designed a sample mixture data augmentation for objects, called PointObjectMix(POM). For datasets with instance labels, we are mixing objects from different scenes into new scenes in order to increase the learning capability of the network while solving the sample balancing problem. The diagram of PointObjectMix is shown in Figure 2.

3.1.2 Feature Extraction Backbone: After PointObjectMix data augmentation, we use Ground Truth supervision for the initial value extraction of the feature extraction network. We use only a very small number of samples to train the initial feature extraction network, and the samples used are not duplicated with the subsequent unsupervised samples to avoid the influence of labels on the unsupervised network. Point cloud feature extraction networks are developing rapidly, and there are many complex networks proposed and used. However, we use the lightweight network DGCNN(Wang et al., 2019) considering the complexity of the method and the subsequent deployment and other related issues. The network extracts features of the local shape of the point cloud by EdgeConv, while still being able to maintain alignment invariance. Also, we add self-attention after EdgeConv to rearrange the features in order to extract global features.

3.2 Self-supervised Channel

In order to reduce the reliance of deep learning networks on data annotation, we designed self-supervised learning channels. In this channel, we introduce super-point-level features to expand the network receptive field and enable the network to learn features at multiple scales. And we use dual scales of point-level and super-point-level contrastive learning strategies to further enhance the accuracy of the feature extraction network. We also designed a corresponding dynamic data augmentation(DDA) module for super-point-level data contrastive learning. The following will describe the components and roles of each module separately in the order of the self-supervised channel.

3.2.1 Super-point Map Generation: Super points are an over-segmentation of point clouds, which can semantically group points of similar geometric features. The super point map can reduce the redundant information of the point cloud, and reduce the cost of subsequent point cloud processing while

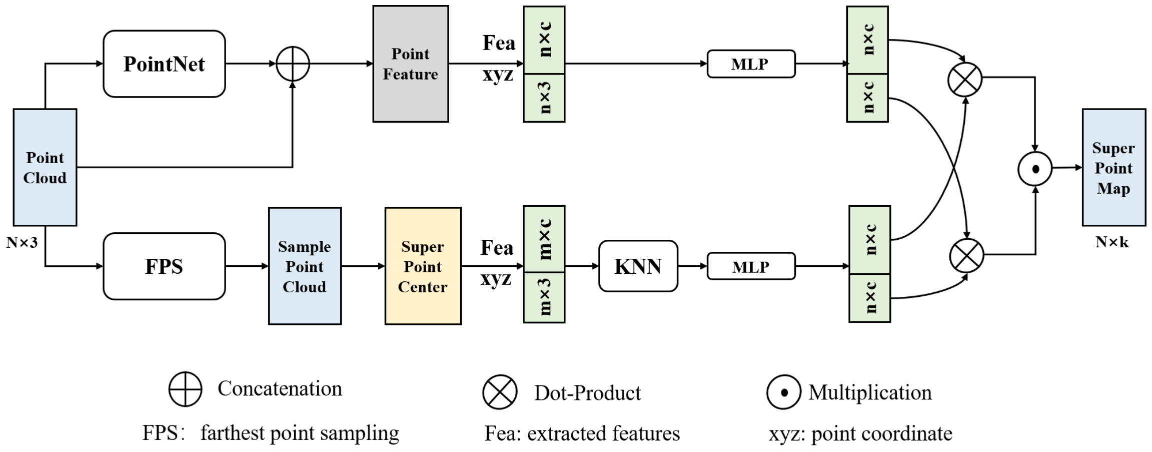

Figure 3. The Overall Architecture of Super Point Generation

aggregating the neighborhood information and expanding the receptive field. Because of its efficient computation and representation, there are already many tasks that use super point maps to represent point clouds, such as 3D detection and semantic segmentation. Also due to the complexity of point cloud data, the generation method of super point map has been investigated. In this work, we follow the super point generation network used in SPNet for super point map generation.

Given the point cloud $P = \{p_i \in R^3 | i = 1, ..., n\}$ with n points, a point-super point association map $H \in Z_n \times m$ between the points and super point centers. We also built a lightweight super point center generation network based on PointNet. We combine the dual-scale features by mapping the point cloud sampled through the farthest point sampling point cloud and the original point cloud to the feature dimension through a weight-sharing PointNet network. We obtain the initial super point centers by feature aggregation. Also for each point in the original point cloud, the association of that point with its closest point is calculated. The association for the i-th point with the j-th super point is calculated as follows:

$$G^t_{ij} = \vartheta^T(p_i, x_j)g(p_i) \bullet \varphi^T(f_i, s_j)h(f_i) \tag{1}$$

$$\theta(p_i, x_j) = RELU(W^T_\theta(p_i - x_j)) \tag{2}$$

$$\varphi(f_i, s_j) = RELU(W^T_\varphi(f_i - s_j)) \tag{3}$$

where $x_j \in R^3$ is the spatial coordinate of the super point center and $s_j \in R^c$ is the feature of the super point center.

The softmax function is also used to calculate the probability that the point belongs to this super point region, so as to obtain the mapping relationship between the point and the super point. The $g(\cdot)$ and $h(\cdot)$ functions are implemented via MLP. W_θ and W_φ are the weights to be learned, and ReLU is the activation function. And we use the difference between the point feature and the center of the super point feature for encoding. The mapping relationship G between the i points of the planning neighborhood and the super points is calculated as follows:

$$\widehat{G^t_{ij}} = \frac{exp(G^T_{ij})}{\sum^k_{l=1} G^T_{il}} \tag{4}$$

And figure 3 shows the overall architecture of super point generation.

3.2.2 Dynamic Data Augmentation: Positive and negative samples are at the core of what makes comparison learning work. Data augmentation is a common method for generating sample pairs in contrastive learning. Inspired by (Li et al., 2020, Li et al., 2022), we propose a dynamic data augmentation module(DDA) for the data organization of super point maps. The method achieves learnable dynamic point cloud data augmentation by MLP and noise signals. We first use PointNet, a lightweight network, for original point cloud P feature extraction. Then Gaussian noise H, G of comparable dimensionality is generated using independent mappings different from the feature extraction. Meanwhile, we plan to use the network simulation affine transformation to map the Gaussian noise G_1 and G_2 to the ordered feature aggregation dimension by MLP to obtain G_t and G_a. Finally, the augmented sample S^{u2} is generated using G_t and G_a, $S^{u2} = G_t \cdot S^u + G_a$. The augmented samples enrich the data diversity in contextual displacement and generate different transformations in the same scene. Figure 4 shows the structure of the dynamic data augmentation module.

3.2.3 Dual-scale Contrastive Learning: We constructed a consistent contrastive strategy learning for both point-level and super-point-level scales. We assume that for two different views of the same object, the features obtained by a robust feature extraction network should be the same. This consistent training allows the network to be robust to low-level feature input perturbations. Also, a stable high-dimensional feature is extracted for the target. Formally, given a point cloud $P^u \in R^{N \times D}$, The super point map S^u is first obtained by the super point generation module. Then our network applies two different groups of data augmentations to create its two views $S^{u1} \in R^{N \times D}$,$S^{u2} \in R^{N \times D}$ respectively. To better convey the point cloud context information as well as to reduce the data processing effort, we use the dynamic data augmentation module described above to complete the data augmentation. Then, the two obtained augmented samples are fed into a weighted backbone network to obtain two high-dimensional features U^1 and U^2 at the super point level. Also, we obtain the recombination feature U^{1A} for one of the high-dimensional features after the self-attention layer in order to increase the effectiveness of the network feature extraction. And then we perform the first stage of contrastive learning for two features U^{1A} and U^2 at the super point level. We back-project the super-point features back to the original point cloud scale U^{1AP} by the mapping relationship between points and super-points. The feature is compared with the high-dimensional feature U^p obtained directly from the

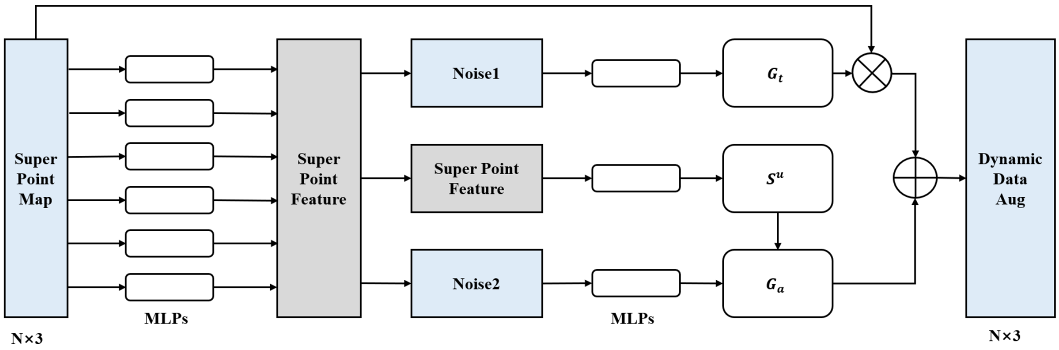

Figure 4. The Structure of Dynamic Data Augmentation Module.

original point cloud by backbone for point-level scale learning.

3.3 Loss Function

We first obtained the super-point level features U^{1A} and U^2 by the feature extraction backbone. We project U^{1A} and U^2 onto an invariant space R^d where the contrastive loss is applied. The goal is to maximize the similarity of U^{1A} and U^2 while minimizing the similarity with all the other projected vectors in the mini-batch of point clouds. We used the NT-Xent loss function in contrastive learning SimCLR(Chen et al., 2020a). NT-Xent loss function is calculated as follows:

$$l(k, U^{1A}, U^2) = -log(\frac{exp\left(s\left(U^{1A}, U^{1A}\right)/\tau\right)}{\sum_{k=1k\neq i}^{N} exp(s(U^{1A}, U^2)/\tau) + \sum_{k=1}^{N} exp(s(U^{1A}, U^2)/\tau)}) \tag{5}$$

where N is the mini-batch size, τ is the temperature co-efficient and s(·) denotes the cosine similarity function. Our super point level instance discrimination loss function L_{sp} for a mini-batch with super point level can be described as:

$$L_{sp} = \frac{1}{2N}\sum_{i=1}^{N}[l(i, U^{1A}, U^2) + l(i, U^2, U^{1A})] \tag{6}$$

At the same time, we project the super point level features to the point level and make the contrastive learning with the features obtained through the original point cloud at the point level. The same NT-Xent loss function is used to train the feature extraction network. In the invariance space, we aim to maximize the similarity of U^P with U^{1AP} since they both correspond to the same objects. Specifically, the point-level loss function is calculated as follows:

$$l(k, U^P, U^{1AP}) = -log(\frac{exp\left(s\left(U^P, U^{1AP}\right)/\tau\right)}{\sum_{k=1k\neq i}^{N} exp(s(U^P, U^{1AP})/\tau) + \sum_{k=1}^{N} exp(s(U^P, U^{1AP})/\tau)}) \tag{7}$$

Our point level instance discrimination loss function L_p for a mini-batch can be described as:

$$L_p = \frac{1}{2N}\sum_{i=1}^{N}[l(i, U^P, U^{1AP}) + l(i, U^{1AP}, U^P)] \tag{8}$$

Finally, we obtain the resultant loss function during training as the combination of L_{sp} and L_P, where L_{sp} represents the super point level feature consistency and L_p represents the point level feature consistency.

$$L = L_{sp} + L_p \tag{9}$$

4. RESULT AND DISCUSSION

4.1 Pretraining Channel

4.1.1 Dataset: We pre-trained the SPDC using less than 10% of the ScannetV2 dataset. The ScannetV2 dataset has a total of 1513 acquisition scenes with 21 categories. There are 1201 scenes in the dataset for training and 312 scenes for testing. We selected 100 scenes point cloud for data network pre-training. During the pre-training process, we sampled the data from each scene in order to train the network end-to-end, so that the number of points in each scene was the same. We use 2048 points for each point cloud. Also in the training phase, we performed POM data augmentation for the dataset.

4.1.2 Implementation Details: To reduce the parameter size of the network and to facilitate comparison with existing methods, we used DGCNN as the feature extractor for the entire network. Also, in order to augment the network's access to the global information of the input scene, we add the self-attention layer after the DGCNN. The dual-scale feature extractor of the entire network is composed of DGCNN + self-attention layer. The Adam optimizer is also used with an initial learning rate of 0.001 and a weight decay of 1×10^{-4}. Cosine annealing is also used to achieve learning rate reduction.

4.2 Segmentation Performance

We evaluate the performance of the SPDC network for the point cloud semantic segmentation task. We use the full S3DIS dataset to test the effectiveness of the network. S3DIS is a large dataset of indoor scenes, containing 271 rooms in a total of 13 categories. This dataset has become a common data benchmark and evaluation metric for point cloud semantic segmentation and instance segmentation. Again, we chose the same parameter settings as in the pre-training phase. The learning rate is 0.001 and a weight decay is 1×10^{-4}. We trained a total of 200 epochs on the complete dataset. Our network mainly trains a feature extractor and uses SVM to act as classifiers in downstream tasks. We achieved 63.3% mIoU in the S3DIS dataset semantic segmentation task through extensive experiments as well as parameter tuning. A comparison of the segmentation results of other methods in the S3DIS dataset is shown in Table 1. The SPDC feature extractor outperforms the state-of-the-art

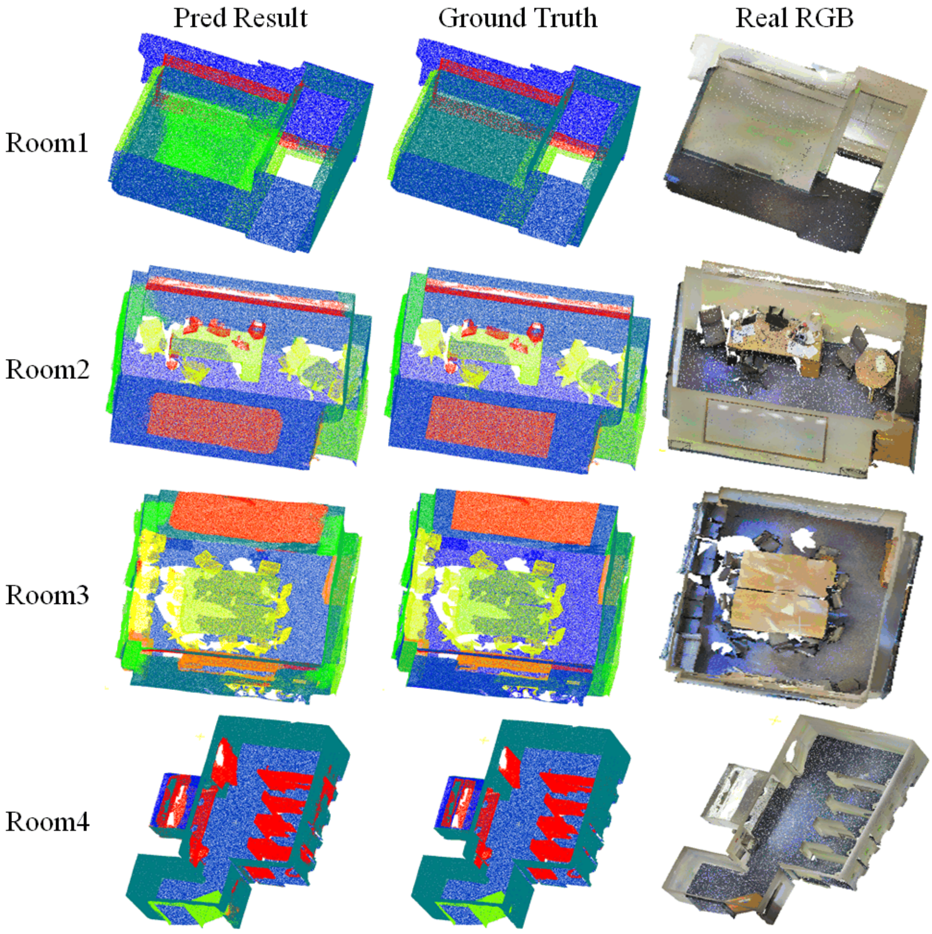

Figure 5. The Visualization Results of Segmentation on S3DIS Dataset.

self-supervised methods available today. In particular, we focus on comparing the more classical point cloud self-supervised method CrossPoint. The method also employs a contrastive learning strategy and also uses DGCNN as a feature extractor. Our method outperforms CrossPoint results by 4.4% in terms of results. However, because of the different training methods and backbone network choices, some methods cannot be fairly compared. Figure 5 shows the segmentation results of our SPDC network.

4.3 Ablation Experiments and Analysis

Our network consists of three main parts, pre-training channels, super-point-assisted feature extraction, and Backbone selection. We performed corresponding ablation experiments in order to verify the usefulness of each module.

4.3.1 Pretraining Channel: To verify the effectiveness of the pre-training channels, we removed the pre-training channels from the framework and kept only the network structure of the dual-scale contrastive learning below. This version is a true departure from the point cloud annotation of the network. The network performs feature extraction entirely through DGCNN, and then dual-scale contrast learning is used as pseudo-supervision of the network. As shown in Table 2, a segmentation accuracy of 51.3% was obtained, indicating that the network is generally effective, but the overall network accuracy is low because it has not been fine-tuned for specific downstream tasks. This also demonstrates the effectiveness of the pre-training channel as a side effect.

Table 1. Comparison of the mean IoU of semantic segmentation results with self-supervised methods on S3DIS.

Method	Supervision	MIoU
PointNet(Qi et al., 2017a)	100%	41.1
PointConv(Wu et al., 2019)	100%	57.3
SPGraph(Landrieu and Simonovsky, 2018)	100%	58.0
MinkowskiNet(Choy et al., 2019)	100%	65.4
KPConv(Thomas et al., 2019)	100%	67.1
DGCNN(Wang et al., 2019)	100%	56.1*
DGCNN+CRF(Xu and Lee, 2020)	0.2%	44.5
MT(Tarvainen and Valpola, 2017)	10%	47.9
DGCNN+CRF(Xu and Lee, 2020)	10%	48.0
MulPro(Su et al., 2022)	10%	49.0
OTOC(Liu et al., 2021)	0.02%	50.1
MIL transformer(Yang et al., 2022)	0.02%	51.4
HybridCR(Li et al., 2022)	0.03%	51.5
GaIA(Lee et al., 2023)	0.02%	53.7
DAT(Wu et al., 2022b)	0.02%	54.6
OTOC++(Liu et al., 2023)	0.02%	56.6
CrossPoint(Afham et al., 2022)	0%	58.4
PointSmile(Li et al., 2023)	0%	58.9
PointMatch(Wu et al., 2022a)	0.1%	63.4
SPDC(No pre-training)	0%	**51.3**
SPDC(No super-point)	0%	**56.7**
SPDC(No self-attention)	0%	**59.6**
SPDC(Completed)	0%	**63.3**

Table 2. Ablation Study of Modules in SPDC.

Pre-training	Super-point	Self-attention	mIoU(%)
✗	✗	✗	51.3
✓	✗	✗	56.7
✓	✓	✗	59.6
✓	✓	✓	**63.3**

4.3.2 Super-point Level Feature: The main innovation of this paper is the introduction of super point level features to assist the semantic segmentation task of point clouds and the design of a corresponding data augmentation module for super point maps. The introduction of the super point map helps the network to better learn the information within the point cloud neighborhood. Also, the over-segmentation of the point cloud indirectly increases the receptive field of the network and makes the network feature extraction more accurate. Meanwhile, in order to verify the effectiveness of the super point module, we did the corresponding ablation experiments. We remove the super point module from the network and just use point-level single-scale features for contrastive learning. As shown in Table 2, the network without super points is able to achieve a segmentation mIoU of 56.7%.

4.3.3 Self-attention Layer: With the widespread use of transformer in computer vision, the attention mechanism has started to be noticed by everyone. Attentional mechanisms are widely used in natural language processing and 2D image processing work for their powerful sequence modeling capabilities. Due to the complexity and disorder of point cloud data, more Transformer-based point cloud processing networks have also been proposed recently. In our network, we use self-attention to complement the lack of global modeling of the scene by the DGCNN feature extractor. To verify the validity of the module, we still chose to remove it for the corresponding ablation experiments. As shown in Table 2, the network without a self-attention layer is able to achieve a segmentation mIoU of 59.6%.

5. CONCLUSION

We propose a dual-scale contrastive learning method called SPDC based on the fusion of super-point and point. The network utilizes point cloud over-segmentation, which is in the form of a super point map to complement the original point cloud feature information. And the network feature extraction capability is trained by contrastive learning on both point-level and super-point-level scales while getting rid of the reliance on data annotation for deep learning networks. Meanwhile, in the contrastive learning process, we designed POM data augmentation patterns for different data structures of the original point cloud and super points, and a learnable dynamic data augmentation module, respectively. Impressively, SPDC achieves SOTA performance among unsupervised networks on the semantic segmentation task of S3DIS datasets. And it shows high robustness after very little fine-tuning.

REFERENCES

Afham, M., Dissanayake, I., Dissanayake, D., Dharmasiri, A., Thilakarathna, K., Rodrigo, R., 2022. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9902–9912.

Ben-Shabat, Y., Avraham, T., Lindenbaum, M., Fischer, A., 2018. Graph based over-segmentation methods for 3D point clouds. *Computer Vision and Image Understanding*, 174, 12–23. https://linkinghub.elsevier.com/retrieve/pii/S107731421830078X.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations. *International conference on machine learning*, PMLR, 1597–1607.

Chen, Y., Hu, V. T., Gavves, E., Mensink, T., Mettes, P., Yang, P., Snoek, C. G., 2020b. Pointmixup: Augmentation for point clouds. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, Springer, 330–345.

Chen, Y., Liu, J., Ni, B., Wang, H., Yang, J., Liu, N., Li, T., Tian, Q., 2021. Shape Self-Correction for Unsupervised Point Cloud Understanding. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 8362–8371.

Choy, C., Gwak, J., Savarese, S., 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3075–3084.

Fan, S., Dong, Q., Zhu, F., Lv, Y., Ye, P., Wang, F.-Y., 2021. SCF-Net: Learning Spatial Contextual Features for Large-Scale Point Cloud Segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Nashville, TN, USA, 14499–14508.

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A., 2020. RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Seattle, WA, USA, 11105–11114.

Huang, S., Xie, Y., Zhu, S.-C., Zhu, Y., 2021. Spatio-temporal Self-Supervised Representation Learning for 3D Point Clouds. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 6515–6525.

Hui, L., Yuan, J., Cheng, M., Xie, J., Zhang, X., Yang, J., 2021. Superpoint Network for Point Cloud Oversegmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 5490–5499.

Landrieu, L., Boussaha, M., 2019. Point Cloud Oversegmentation With Graph-Structured Deep Metric Learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Long Beach, CA, USA, 7432–7441.

Landrieu, L., Simonovsky, M., 2018. Large-scale point cloud semantic segmentation with superpoint graphs. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4558–4567.

Lee, M. S., Yang, S. W., Han, S. W., 2023. Gaia: Graphical information gain based attention network for weakly supervised point cloud semantic segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 582–591.

Li, M., Xie, Y., Shen, Y., Ke, B., Qiao, R., Ren, B., Lin, S., Ma, L., 2022. Hybridcr: Weakly-supervised 3d point cloud semantic segmentation via hybrid contrastive regularization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14930–14939.

Li, R., Li, X., Heng, P.-A., Fu, C.-W., 2020. Pointaugment: an auto-augmentation framework for point cloud classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6378–6387.

Li, X., Wei, M., Chen, S., 2023. PointSmile: Point Self-supervised Learning via Curriculum Mutual Information. *arXiv preprint arXiv:2301.12744*.

Liu, Z., Qi, X., Fu, C.-W., 2021. One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1726–1736.

Liu, Z., Qi, X., Fu, C.-W., 2023. One Thing One Click++: Self-Training for Weakly Supervised 3D Scene Understanding. *arXiv preprint arXiv:2303.14727*.

Liu, Z., Tang, H., Lin, Y., Han, S., 2019. Point-Voxel CNN for Efficient 3D Deep Learning.

Meng, H.-Y., Gao, L., Lai, Y.-K., Manocha, D., 2019. VV-Net: Voxel VAE Net With Group Convolutions for Point Cloud Segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Seoul, Korea (South), 8499–8507.

Milioto, A., Vizzo, I., Behley, J., Stachniss, C., 2019. RangeNet ++: Fast and Accurate LiDAR Semantic Segmentation. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, Macau, China, 4213–4220.

Papon, J., Abramov, A., Schoeler, M., Worgotter, F., 2013. Voxel Cloud Connectivity Segmentation - Supervoxels for Point Clouds. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Portland, OR, USA, 2027–2034.

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Qi, C. R., Yi, L., Su, H., Guibas, L. J., 2017b. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space.

Sha, Z., Chen, Y., Li, W., Wang, C., Nurunnabi, A., Li, J., 2020. A BOUNDARY-ENHANCED SUPERVOXEL METHOD FOR EXTRACTION OF ROAD EDGES IN MLS POINT CLOUDS. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B1-2020, 65–71. https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLIII-B1-2020/65/2020/.

Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E., 2015. Multi-view Convolutional Neural Networks for 3D Shape Recognition. *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Santiago, Chile, 945–953.

Su, Y., Xu, X., Jia, K., 2022. Weakly Supervised 3D Point Cloud Segmentation via Multi-Prototype Learning. *arXiv preprint arXiv:2205.03137*.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L. J., 2019. Kpconv: Flexible and deformable convolution for point clouds. *Proceedings of the IEEE/CVF international conference on computer vision*, 6411–6420.

Wang, H., Liu, Q., Yue, X., Lasenby, J., Kusner, M. J., 2021. Unsupervised Point Cloud Pre-training via Occlusion Completion. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 9762–9772.

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

Wu, W., Qi, Z., Fuxin, L., 2019. Pointconv: Deep convolutional networks on 3d point clouds. *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 9621–9630.

Wu, Y., Yan, Z., Cai, S., Li, G., Yu, Y., Han, X., Cui, S., 2022a. Pointmatch: a consistency training framework for weakly supervisedsemantic segmentation of 3d point clouds. *arXiv preprint arXiv:2202.10705*.

Wu, Z., Wu, Y., Lin, G., Cai, J., Qian, C., 2022b. Dual adaptive transformations for weakly supervised point cloud segmentation. *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, Springer, 78–96.

Xie, S., Gu, J., Guo, D., Qi, C. R., Guibas, L. J., Litany, O., 2020. PointContrast: Unsupervised Pre-training for 3D Point Cloud Understanding.

Xu, X., Lee, G. H., 2020. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13706–13715.

Yang, C.-K., Wu, J.-J., Chen, K.-S., Chuang, Y.-Y., Lin, Y.-Y., 2022. An mil-derived transformer for weakly supervised point cloud segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11830–11839.

Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J., 2022. Point-BERT: Pre-training 3D Point Cloud Transformers with Masked Point Modeling. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, New Orleans, LA, USA, 19291–19300.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., Yoo, Y., 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.

Zhang, H., Cisse, M., Dauphin, Y. N., Lopez-Paz, D., 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, J., Chen, L., Ouyang, B., Liu, B., Zhu, J., Chen, Y., Meng, Y., Wu, D., 2022. Pointcutmix: Regularization strategy for point cloud classification. *Neurocomputing*, 505, 58–67.