Computers & Graphics 116 (2023) 46–54

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on CAD/Graphics 2023

Weakly supervised semantic segmentation for point cloud based on view-based adversarial training and self-attention fusion

Yongwei Miao, Guoxiang Ren, Jinrong Wang, Fuchang Liu *

School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China
Key Lab. of Cryptography of Zhejiang Province, Hangzhou Normal University, Hangzhou 311121, China

ARTICLE INFO

Article history:
Received 28 April 2023
Received in revised form 11 July 2023
Accepted 5 August 2023
Available online 9 August 2023

Keywords:
Point cloud
Semantic segmentation
Weakly supervised learning
Self-attention mechanism
Siamese loss
Smoothness loss

ABSTRACT

Traditional methods of weakly supervised semantic segmentation (WSsegmentation) for point cloud scenes have several limitations including limited precision and difficulty in handling complex scenes due to imprecise labels or partial annotations. To address these issues, we perform view-based adversarial training on the original point cloud scene samples through view resampling and Gaussian noise perturbation to reduce overfitting. Combining a self-attention mechanism with multi-layer perceptrons and point cloud segmentation strategy, we can perform dimensionality enhancement and dimensionality reduction operations to better capture the local features of point cloud data. Finally, we can obtain the semantic segmentation results of point cloud scene by fusing local and global semantic features. In the design of the network loss function, we combine the Siamese loss and the smoothness loss and the cross-entropy loss to improve the ability and fidelity of semantic networks. Specifically, the Siamese loss is used to compute the distance between different augmented point cloud data in their feature embedding space and the smoothness loss is used to penalize the discontinuity of semantic information between adjacent regions. The proposed weakly supervised segmentation network achieves an overall segmentation accuracy close to fully supervised segmentation methods and outperforms most of the existing weakly supervised segmentation methods by 5% to 10% in scene segmentation in terms of mIoU on S3DIS, ShapeNet, and PartNet datasets. Extensive experiments demonstrate the robustness, effectiveness, and generalization of the proposed point cloud segmentation network.

1. Introduction

Point cloud semantic segmentation is a cornerstone problem in computer graphics and 3D vision. Scene point cloud data has been widely used due to its advantages such as convenient 3d scanning [1] and effective representation of complex scenes [2]. Most of fully supervised point cloud semantic segmentation methods usually require daunting task of supervised semantic labeling during scene understanding and neural network training. The scene label information can effectively guide the segmentation and understanding of point cloud scenes in supervised learning. However, for large-scale indoor scene data with usually millions of discrete sampling points, accurately annotating semantic labels for different objects or parts in the point cloud scene is a highly challenging task and the labor-intensive work. Therefore, it is promising to consider weakly supervised learning strategies that can effectively learn scene segmentation networks with a limited amount of annotation information in the point cloud scene. Unlike traditional point cloud semantic segmentation strategies based on supervised learning, weakly supervised methods only use 1% to 10% of label information. By utilizing a small amount of semantic labeling information, the WSsegmentation strategy can automatically explore semantic features and labels of point cloud data, thus achieving efficient scene semantic segmentation and greatly reducing data annotation costs. The techniques of weakly supervised segmentation have already been used in autonomous driving and indoor navigation tasks [2]. Therefore, the weakly supervised segmentation has been a highly promising research area that has attracted significant attention.

Despite the development of WSsegmentation methods, they still cannot achieve comparable level of accuracy and effectiveness as fully supervised methods due to limitations such as insufficient data annotation. Especially in practical applications, a trade-off needs to be made between data annotation costs and model performance. Furthermore, the WSsegmentation strategies have certain limitations in dealing with complex point cloud

* Corresponding author at: School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China.
E-mail addresses: ywmiao@hznu.edu.cn (Y. Miao), liufc@hznu.edu.cn (F. Liu).

https://doi.org/10.1016/j.cag.2023.08.007
0097-8493/

scenes. For instance, in scenarios where the scene contains occlusions, overlaps, or reflections, weakly supervised models may produce incorrect segmentation results. Moreover, when dealing with partially occluded or missing data, inaccurate segmentation results may occur on accurately segmenting the boundaries between objects of different shapes, different shape scales, and different semantic categories in the scene. To overcome these limitations, this paper proposes a WSsegmentation network that simultaneously introduces Gaussian noise perturbation and view resampling to enhance point cloud data via effectively augmenting point cloud data information, thereby helping to improve the robustness and generalization of the segmentation network. Specifically, the combination of self-attention modules and multilayer perceptrons is used in the network to strengthen the extraction of local features of point cloud and improves the semantic segmentation accuracy of local structures of point cloud in the scene. In addition, this paper uses the Siamese loss and the smoothness loss to compute the distance between different augmented point clouds in their feature embedding space and penalize the color discontinuity between adjacent point cloud segmentation regions. We also use the cross-entropy loss as part of the final loss between a small amount of point cloud data labels and ground truth to improve the accuracy of point cloud segmentation in the complex scenes.

The main technical contributions of this work can be summarized as follows:

- We propose effective data augmentation methods, including view resampling and adversarial training, to tackle partial annotations;
- To fully integrate the global and local features of point cloud, we employ high-dimensional feature self-attention mechanism and multi-layer perceptron structure to fuse multi-level features, and leverage residual connections to effectively reduce the gradient vanishing issue during neural network training;
- We present a novel point cloud segmentation framework, which can achieve the higher segmentation accuracy than traditional segmentation networks for different point cloud scenes.

2. Related works

In this section, we introduce relevant works on point cloud data augmentation and weakly supervised point cloud segmentation.

2.1. Data augmentation for weakly supervised learning of point cloud

Point cloud data augmentation in weakly supervised learning usually includes random rotation, model scaling, and color jittering [3]. These augmentation techniques effectively improve the generalization ability of neural network training. Kim et al. [4] used locally weighted transformations to produce non-rigid deformations with a kernel density-based estimation method that applies the weights being inversely proportional to the distance between sample points and the target point. Based on an adversarial training strategy, Li et al. [5] used Gaussian noise and masking techniques to automatically generate augmented point cloud. This method filters Gaussian noise using randomly generated binary masks to enhance features and reduce noise effects. However, these methods mainly focus on discrete point cloud and cannot simulate the effects of different viewpoint changes as well as different point cloud noises that may exist in point cloud data acquisition in real applications, which is difficult for neural networks to achieve ideal semantic segmentation results during training. In this paper, the point cloud semantic segmentation network introduces view resampling and Gaussian noise perturbation simultaneously to augment point cloud data and generate diverse point cloud samples for training segmentation networks.

2.2. Point cloud semantic segmentation and weakly supervised learning

Semantic segmentation strategies based on neural networks have shown great advantages in semantic segmentation tasks [6–8]. Some existing works convert point clouds into regular grid data based on views or volumetric representations [9]. Hou et al. [10] jointly learn image features and three-dimensional geometric features using a segmentation network to achieve three-dimensional instance segmentation. However, methods based on multi-view image data representation have certain limitations due to object occlusion and complex data input in the scene, and voxel representation is greatly restricted due to its high storage consumption. Nan et al. [11] propose a search-based classification and segmentation method for scene modeling with the assistance of pre-trained object categories. Li et al. [12] propose an object retrieval approach to replace scanned 3D data with objects from shape database and recover the scene based on them. The method proposed by Mattausch et al. [13] detects repeated objects in the scene by detecting scan data in multiple rooms. Wu et al. [14] propose a consistency training framework, which generates pseudo-labels by randomly perturbing and rotating unlabeled point cloud, and applies these pseudo-labels to the training of weakly supervised models to improve the robustness and generalization ability of the model. Owing to the definition of neighborhood co-occurrence matrix (NCM) for point cloud data, Gong et al. [7] proposed a method for generating target NCM and prediction NCM from semantic labels, which can help to indoor scene segmentation. Zhao et al. [8] proposed a segmentation network LG-Net to efficiently learn the local dual features and global correlations for large-scale point clouds. The attention mechanism and transformer structure can also be employed for improving the performance of point cloud semantic segmentation [15,16]. Hu et al. [17] proposed a WSsegmentation method which is suitable to large-scale discrete point cloud data by training a point cloud classifier to achieve effective semantic segmentation. The WSsegmentation network proposed in this paper fully utilizes data augmentation strategies to improve the generalization and robustness of weakly supervised segmentation.

3. Proposed point cloud weakly supervised segmentation network

3.1. Overview

In this paper, we propose a WSsegmentation network based on data augmentation and self-attention mechanism. The overall pipeline of the proposed network is illustrated in Fig. 2, which first performs view-based adversarial training via data augmentation on the original point cloud data by view resampling and Gaussian noise perturbation, and then uses a point cloud feature extraction module composed of self-attention model and multilayer perceptron to respectively extract and aggregate features of the augmented point cloud data, and finally obtains the point cloud segmentation results through max pooling layer and multilayer perceptron dimensionality reduction. Meanwhile, a joint loss functions are introduced into the network to improve the accuracy of point cloud semantic segmentation.

(a) (b)

Fig. 1. Data augmentation result (b) for the input office point cloud scene (a).

3.2. View-based adversarial training

In order to effectively improve the effectiveness and accuracy of point cloud scene semantic segmentation, the size of the scene dataset is usually an important factor in training neural network models. However, the limited amount of real scene point cloud data greatly affects the generalization ability of neural network models, so it is necessary to effectively augment the point cloud scene data [3]. Scene point cloud data augmentation can be achieved by performing a series of transformations on the original point cloud scene, such as rotation [4] or adding Gaussian noise [5] to generate new scene point cloud data and expand the scene point cloud dataset, ultimately increasing the number of sample data required for model training. In the point cloud data augmentation process, performing both viewpoint transformation and adding Gaussian noise can complement each other and provide augmented data for adversarial training. Specifically, we first perform viewpoint transformation on the original point cloud and then employ data augmentation by adding Gaussian noise to ensure that the noise can exist in one point cloud scene with different views, thereby increasing the data diversity for effective network training.

3.2.1. View resampling

To increase the diversity of point cloud data and improve the model's generalization ability, view resampling is applied to scene point cloud data based on different viewing directions or observation angles to generate point cloud data from different perspectives. View resampling helps segmentation models better recognize and distinguish different objects in the scene, and increases the number of samples of scene point cloud data, thereby increasing the amount of sample data for model training. Specifically, a training sample $\tilde{X}$ is first given, and then a random transformation R is applied to it, which includes trigonometric functions and several variables that follow specific distributions such as:

$$\boldsymbol{R} = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} (2a-1)c & (2b-1)(1-c) & 0 \\ (2a-1)(1-c) & (2b-1)c & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1}$$

Here, Θ follows a uniform distribution $U(0, 2\pi)$, while a, b, and c follow a Bernoulli distribution $B(1, 0.5)$. The first transformation matrix controls the degree of rotation of the original point cloud data, and the second transformation matrix controls the amplitude of the coordinate transformation of the point cloud mirroring with the original point cloud data. By multiplying the original point cloud data X by the transpose of the point cloud rotation matrix R, the augmented point cloud data sample $X = XR^T$ is obtained. The data enhancement results are shown in Fig. 1.

3.2.2. Gaussian perturbation

For the task of WSsegmentation, the insufficient sample size of scene point cloud data and high labeling costs often result in insufficient training sample data during the neural network training process, leading to model overfitting. The trained network model usually performs well on the training data, but not on the testing data. To alleviate this problem, this paper considers adding Gaussian noise to point cloud as data augmentation after scene point cloud data rotation transformation to further enhance the diversity of scene sample data. Adding Gaussian noise to scene point cloud data can improve the robustness of the trained network model to data noise, increase the diversity of scene data sets, and effectively improve the model's generalization performance.

3.3. Feature extraction and aggregation based on self-attention mechanism

Recently, neural network models typically use self-attention mechanisms and multilayer perceptrons for feature extraction [18]. As an effective form of point cloud data feature representation, the self-attention mechanism is able to focus on the feature information of different position-sampled points in the scene point cloud, and simultaneously learn the local intrinsic relationships between each sampled point and other sampled points in the point cloud data, thus further capturing the global context information in the scene point cloud data. As a form of non-linear transformation, the multilayer perceptron is able to map point cloud data features to a higher dimensional feature space, which helps to obtain an effective representation of high-dimensional features. By stacking multilayer perceptrons, the neural network is able to progressively extract more abstract point cloud high-level feature information, thereby enhancing the model's representation ability. This paper effectively combines the self-attention mechanism and multilayer perceptrons to better capture the global and local features of scene point cloud data, thus improving the accuracy and robustness of scene point cloud semantic segmentation.

Compared to convolutional methods, the self-attention module has stronger capabilities in extracting local features and global structures of point cloud, making it better suited for understanding different types of point cloud scenes. In the fusion process of local and global features of point cloud, this article concatenates point cloud information of different scales in the channel dimension, and inputs it into a neural network for semantic segmentation.

During the training process of a neural network, problems such as exploding or vanishing gradients may arise, making the model difficult to train. As shown in Fig. 3, this article's network architecture uses residual links to allow gradients to be propagated before skipping layers, allowing for effective fusion of input and output features and making the network model easy to train and accelerate convergence. For scene point cloud data, after two rounds of residual links and feature extraction and aggregation using self-attention models, using a multilayer perceptron for dimensionality expansion and fusion operations can obtain an $N*$ 2048 dimensional point cloud feature matrix. Finally, through the max pooling layer and multilayer perceptron dimensionality reduction processing, k point cloud label probabilities represented in logarithmic form are obtained.

4. Loss function

To design of the loss function for the weakly supervised segmentation, we introduce a joint loss to train the proposed point cloud segmentation network.

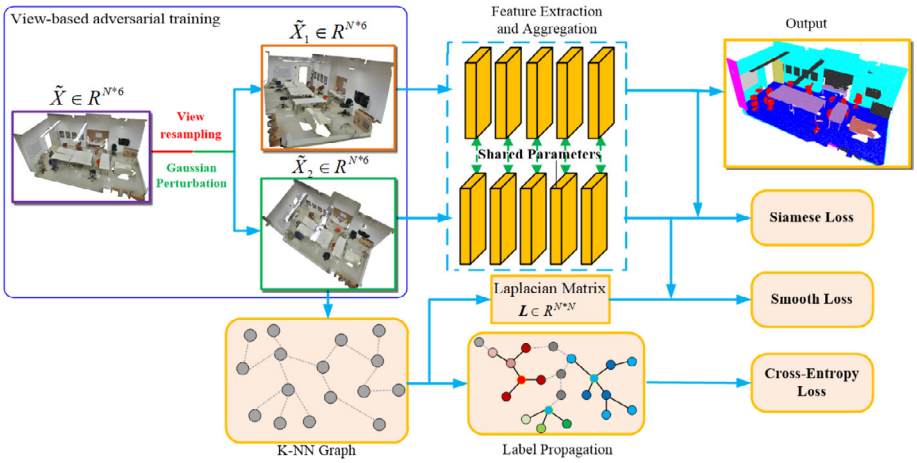


Fig. 2. The pipeline of the proposed WSsegmentation network. Given an input point cloud with partial annotations, our network outputs semantic segmentation results using adversarial training and self-attention based feature extraction and fusion. The detailed structure of feature extraction and aggregation module is given in Fig. 3.

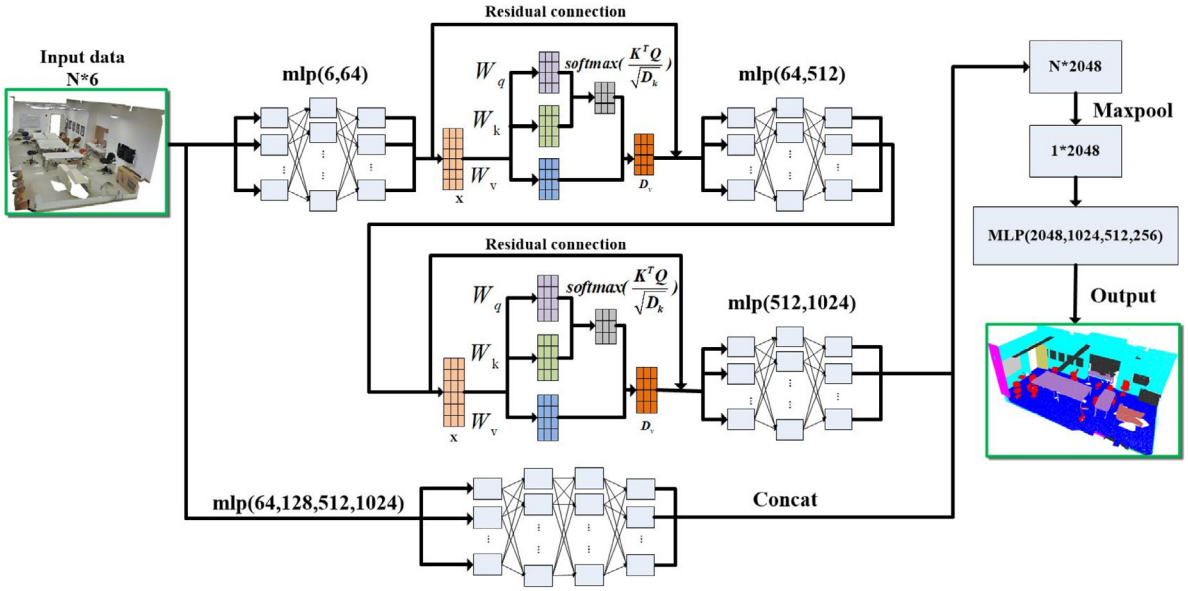


Fig. 3. Feature extraction and aggregation module. Three linear transformations are applied to each feature vector of the sampled points to obtain query, key, and value feature vectors, respectively. The similarity is normalized using softmax to obtain a weight distribution for each sampled point with respect to other sampled points. The weighted average of the value feature vector is computed based on the weight distribution resulting in an aggregated feature vector for each sampled point.

4.1. Siamese loss

The purpose of introducing the Siamese loss is to make the segmentation network model invariant to rotation, translation, and scaling transformations of the point cloud data, thereby improving the generalization performance of the segmentation network model.

Specifically, the Siamese loss is computed as follows: for input point cloud data pairs $P_1 = \{(x_i^1, y_i^1, z_i^1),\ i = 1, 2, \ldots, N\}$ and $P_2 = \{(x_i^2, y_i^2, z_i^2),\ i = 1, 2, \ldots, N\}$, with feature vectors represented as $f(\mathbf{P}_1)$ and $f(\mathbf{P}_2)$, respectively, the distance or similarity between the two feature vectors can be calculated using methods such as Euclidean distance or cosine distance. A commonly used method is to use the following Euclidean distance calculation:

$$dist = \|f(\mathbf{P}_1) - f(\mathbf{P}_2)\|^2 \tag{2}$$

Here it is necessary to convert the distance between feature vectors into similarity. This can generally be achieved by using the softmax function $g(\cdot)$. Specifically, $g(f(\mathbf{P}_1))$ and $g(f(\mathbf{P}_2))$ represent the probability sizes of the label prediction for the sampling points by the weight-sharing feature extraction and aggregation module. In this article, the Siamese loss [19] is measured using Euclidean distance as follows:

$$L_{siamese} = \frac{1}{B \cdot N \cdot K} \sum_b \|g(f(\mathbf{P}_1^b)) - g(f(\mathbf{P}_2^b))\|_F^2 \tag{3}$$

where B represents the logarithm of the input point cloud scene for each batch when calculating the loss, and each point cloud scene is composed of N discrete sample points. The feature dimension of each sample point is K, b represents the index of the point cloud data, and F represents the vector norm.

4.2. Smoothing loss

Smoothing loss is a type of smoothness-constrained loss function that is typically used to penalize the prediction differences

between adjacent points in a point cloud. In point cloud semantic segmentation, the smoothing loss compares the label prediction values of each sample point with those of its neighboring sample points, and calculates the prediction error between the two sampling points, thereby improving the segmentation accuracy of neural network models.

To construct the manifold graph of the point cloud scene, the pairwise distance matrix $\mathbf{P}_c = (P_{ij}^c) \in R^{N*N}$ (where $P_{ij}^c = \| P_i^c - P_j^c \|_2$) is first calculated for the coordinate information channel or color information channel c. Then, by searching for the k nearest neighboring sample points $N_k(\mathbf{p})$ for each sample point, the k-NN graph can be constructed, and the corresponding weight matrix $W^c \in R^{N*N}$ is written as:

$$w_{ij}^c = \begin{cases} \exp(-P_{ij}^c/\eta), \mathbf{p}_j \in N_k(\mathbf{p}_i) \\ 0, otherwise \end{cases}, \ i, j \in \{1, \cdots N\}. \tag{4}$$

If both the coordinate information in the *xyz* channel and the color information in the *rgb* channel are available, this article uses the sum of two weight matrices to generate a more reliable point cloud scene manifold graph, i.e., $w_{ij}^c = w_{ij}^{xyz} + w_{ij}^{rgb}$. This is because using only the coordinate information in the *xyz* channel can lead to the neglect of color information, thus reducing the segmentation accuracy of objects with similar shapes but different colors. On the other hand, using only the color information in the *rgb* channel may result in distant samples of different categories being misclassified as the same label. Here, we introduce the following smoothness loss function for network training:

$$L_{smooth} = \frac{1}{\|W^c\|_0} \sum_i \sum_j w_{ij}^c \| f(X_i) - f(X_j) \|_2^2 \tag{5}$$

Specifically, w_{ij} represents the similarity of labels between sampling points. When w_{ij} is larger, the predicted values between different sampling points are closer. $f(X_i)$ and $f(X_j)$ represent the predicted values between adjacent sampling points.

4.3. Cross-entropy loss

The purpose of introducing cross-entropy loss is to optimize and update the parameters of the point cloud segmentation model, and improve the performance of the segmentation network. The calculation of the cross-entropy loss is as follows:

$$L_{cross} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} y_{i,j} log(p_{i,j}) \tag{6}$$

where N represents the number of sampling points, K represents the number of labels, y_{ij} represents the probability of the ith sampling point's true label j, and p_i, j represents the probability of the ith sampling point's predicted label j. By minimizing this loss function, the predicted labels can be made closer to the true labels, thereby improving the accuracy of point cloud semantic segmentation. For the existing methods of WSsegmentation, training data only provides label information for few sampling points. Label propagation is used to transmit label information to unknown points and improve the generalization ability of the experimental model. Specifically, in this paper, the network uses a distance-weighted method to attach a distance-related weight to each sampling point, and obtains the label of an unknown sampling point through distance-weighted averaging using the label information of surrounding known sampling points. These labels are considered pseudo-labels, and in each training process, the neural network updates the probability of each point based on the pseudo-labels, and corrects the pseudo-labels based on the updated probabilities. Through iterative training, more accurate label results are obtained.

4.4. Overall loss and network training algorithm

The final loss function is defined as follows,

$$L = \alpha L_{siamese} + \beta L_{smooth} + \gamma L_{cross} \tag{7}$$

the weight of the loss function can be initially estimated by referring to the data training quantity of different loss functions.

Due to the limited number of labeled points, the cross-entropy loss function only trains and optimizes about 20% of the sampled points in the experiment. Therefore, the initial value of γ is set to 0.2 in this experiment. Meanwhile, since the smoothness loss function and the Siamese loss function have the same amount of training data, their initial values are both set to 0.4. An error value of 0.1 is set for the coefficients of the three loss functions, so that the network in this paper can explore the most accurate segmentation results under different coefficients through training. The specific training process is shown in Algorithm 1. To effectively train the point cloud segmentation network in this paper, the $L_{siamese}$ loss is used to train the segmentation network for 100 epochs in the experiment. Then, the total loss L_{total} is used for additional 100 epochs of training. Otherwise, an additional 100 epochs of training on the total loss function are conducted. In addition, the weights of the loss function are determined to be $\alpha = 0.5$, $\beta = 0.3$.

Algorithm 1: Training algorithm for our network
Input: point cloud $X \in R^{N*6}$, labels $y \in Z^N$, α=0.4, β=0.4, γ=0.2.

Sample $\varphi \sim U(0, 2\pi)$ and $a, b, c \sim B(1, 0.5)$;
Generate augmented sample $X_1 = XR_1{}^T$ and $\tilde{X}_1 = X_1 + \sigma_1 * \xi_1$;
Generate augmented sample $X_2 = XR_2{}^T$ and $\tilde{X}_2 = X_2 + \sigma_2 * \xi_2$;
for *epoch = 1* ***to n*** **do**
    $\theta = \theta - \alpha \nabla L_{siamese}(X_1, X_2)$;
Get the optimal solution: α =0.5; β =0.3; γ=0.2;
for *epoch = 1* ***to n*** **do**
    Generate Graph using K-NN;
    Construct Laplacian L;
    $\theta = \theta - \alpha \nabla L_{total}$;

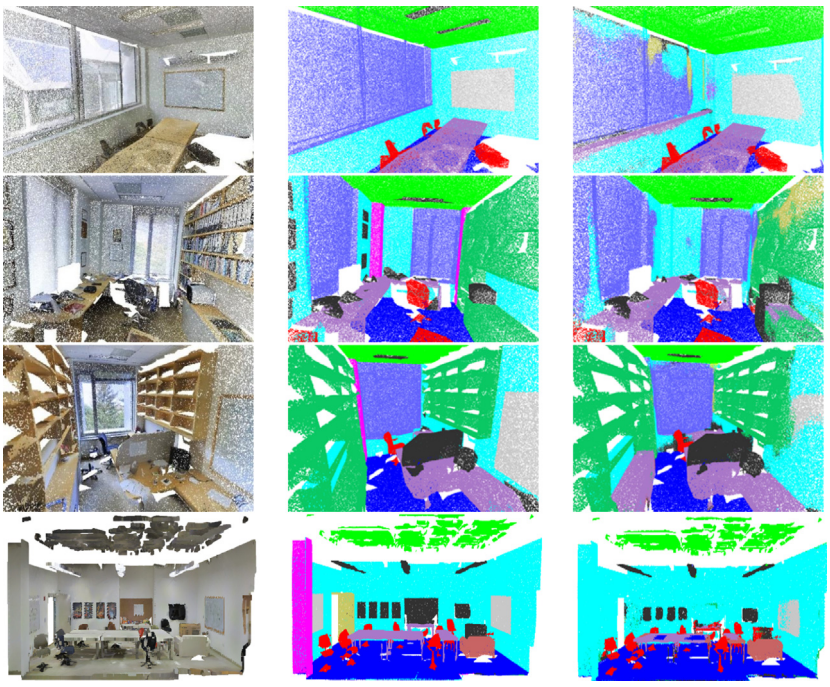
5. Experimental results and discussion

All the experiments have been implemented on the Ubuntu 16.04 system. The hardware environment of the program includes an Intel Core i9-10900K CPU of 3.70 GHz, 500 GB memory, and an NVIDIA GTX 1080 GPU. The software development platform used is Python 3.6 and PyTorch 1.6.

5.1. Dataset and preprocessing

The S3DIS dataset [20] consists of point cloud scene data from 6 different indoor areas, containing over 400 million sampled points. The training set and test set are divided according to different types of areas. Specifically, the first 5 areas in the dataset are usually used for training, while another area (Area 5) is used for testing. This partitioning method ensures that the test set and training set are independent of each other, and also provides a more challenging testing environment.

ShapeNet dataset [21] collects 16 common object categories, each with 16,881 shapes. In the ShapeNet dataset experiments, 70% of the scene point cloud data for each category is usually allocated to the training set, and the remaining 30% of the scene



(a) Raw point cloud data (b) GroundTruth (c) Segmentation results

Fig. 4. The semantic segmentation results in different cloud scenes.

point cloud data is used as the test set. PartNet dataset [22] consists of 24 unique shape categories with a total of 26,671 shapes, such as furniture, cars, animals, tools, electronics, buildings. In this dataset, 80% of the models are selected for the training set, and the remaining 20% models are used for testing.

During training process, we use view-based adversarial training to augment the training data with adversarial examples via view-resampling and Gaussian noise perturbation.

5.2. Segmentation results of our proposed network

Using different labeling strategies, the neural network is evaluated on the ShapeNet dataset [21] for segmentation performance. Table 1 presents the comparison results of different labeling strategies on the ShapeNet dataset for point cloud segmentation, where the labeling strategy is described by $x\%$ (Sampling percentage is abbreviated as $\#S$) representing the percentage of experimental points in the total sampling points of the point cloud model, each sample has $y\%$ of labeled points (abbreviated as $\#P$), and satisfies the fixed label constraint of $x \times y = K$, making the total number of labeled sampling points constant, where $K = 1000$ in the experiments. Five different strategy combinations are evaluated, and the class-average accuracy (mIoU:CatAvg) and sample-average accuracy (mIoU:SampAvg) of point cloud segmentation under different strategies are given. We observe that the scene segmentation accuracy (mIoU) will continue to improve as $x\%$ increases from 10% to 100% under fixed label constraints, labeling more samples extensively with fewer labeled points per sample usually results in more accurate segmentation performance for neural networks compared to labeling more points on fewer samples.

To verify the impact of labeled sampling point ratios on the segmentation results, the network training samples were fixed at 100%, and the effect of different labeling strategies on various datasets with regard to the percentage of labeled sample points are given in Table 2. As the percentage of labeled sample points in the ShapeNet34 dataset [21] decreased from 100% to 10%, the accuracy of point cloud segmentation, mIoU, decreased from 83.32% to 77.42%, a decrease of 5.90%. Similarly, for the PartNet16 dataset [22] and S3DIS dataset [20], the segmentation accuracy, mIoU, decreased by 12.77% and 1.80%, respectively. The experimental results demonstrate that the proposed segmentation network has less dependence on the number of labeled points.

Table 1
Comparison of different labeling strategies on ShapeNet [21] point cloud segmentation.

Label strategy	mIoU: Cat	mIoU: Samp
$\#S = 10\%$, $\#P = 100\%$	70.69	77.32
$\#S = 20\%$, $\#P = 50\%$	72.3	79.25
$\#S = 50\%$, $\#P = 20\%$	75.12	79.45
$\#S = 80\%$, $\#P = 12.5\%$	76.32	80.69
$\#S = 100\%$, $\#P = 10\%$	77.42	81.19

Fig. 4 shows the semantic segmentation results of different indoor point cloud scenes in Area 5 of the S3DIS dataset [20]. It can be seen that the segmentation of different scenes in this dataset is challenging, for example, the window and ceiling frame in the first row of the conference room scene, the bookshelf and cabinet in the second row of the office scene, the blackboard and window frame in the third row of the office scene, the door frame in the fourth row of the corridor scene, and the table, blackboard, and door in the fifth row of the office scene. The lack of sufficient coordinate and color feature information for these structures will result in a decrease in the prediction accuracy of the segmentation network. Nevertheless, the neural network proposed in this article still achieves accurate segmentation results in most indoor point cloud scenes, such as chairs, walls, ceilings, floors, windows, and sofas.

Table 2
Results of different labeling strategies on mIoU of segmentation accuracy of different data sets with 100% sampling.

Label strategy	mIoU (%)		
	ShapeNet [21]	PartNet [22]	S3DIS [20]
$\#P = 100\%$	83.32	67.66	49.12
$\#P = 50\%$	78.11	59.51	48.55
$\#P = 10\%$	77.42	54.89	47.32

5.3. Comparison of different segmentation methods

To compare the segmentation performance of our presented network with other methods, as shown in Table 3, we selected 9 different segmentation methods (such as PointNet++ [23], DGCNN [24], SegCloud [6], PointCNN [25], k-means [26], MT [27], Xu et al. [28], Π Model [29] and MulPro [30]) to compare their segmentation results on the point cloud scenes in Area 5 of S3DIS dataset [20]. The average intersection-over-union (mIoU) of scene segmentation was used as the performance indicator, and the segmentation performance of each method on 12 different object categories was compared.

It can be seen from Table 3 that our proposed segmentation network has higher segmentation accuracy compared with traditional weakly supervised MT method [27], Xu et al. method [28], Π Model method [29], and unsupervised k-means method [27]. The performance of our segmentation method is close to that of fully supervised methods (such as DGCNN network [24], SegCloud network [6]), and weakly supervised MulPro network [30].

Specifically, our network achieves accurate segmentation on seven categories such as ceiling, wall, window, floor, chair, table, and sofa, with segmentation accuracies of 89.98%, 75.80%, 50.10%, 98.20%, 68.90%, 72.10%, and 53.10%, respectively. For the training time of the segmentation network, we trained our network on 204 point cloud scenes in Area 5 of the S3DIS dataset [20], with a total of 195 million points. The traditional PointNet++ network [23] and DGCNN network [24] took 5 h and 13 h to train, respectively, while our network only took 4.5 h, demonstrating its high efficiency. Experimental results show that our network generally has higher segmentation accuracy than other existing networks in the tasks of WSsegmentation.

Fig. 5 shows the comparison of segmentation results on indoor point cloud scenes in Area 5 of the S3DIS dataset [20] using PointNet++ network [23], MT method [28], and our proposed network. It can be observed that the fully supervised segmentation method, PointNet++ network [23], has more accurate segmentation results than the weakly supervised scheme. For example, it performs better than the weakly supervised MT method [27] in segmenting scene boundaries, blackboards, and desktop clutter (shown in yellow ellipses of Fig. 5). The proposed method is between the fully supervised PointNet++ network [23] and the weakly supervised MT method [27] in terms of segmentation results for room doors and windows, scene boundaries, and desktop objects, and is closer to the fully supervised method.

5.4. Ablation experiment

In order to verify the effectiveness of the data augmentation strategy, residual connection module, Siamese loss, smoothness loss and cross-entropy loss proposed in this paper on point cloud scene segmentation, ablation experiments were conducted on the S3DIS dataset [20] and ShapeNet dataset [21]. Table 4 shows the segmentation accuracy (mIoU) of point cloud semantic segmentation under different component selections. It can be seen that the Siamese loss has a significant improvement on the semantic segmentation results of both S3DIS dataset [20] and ShapeNet dataset [21]. As the feature extraction and aggregation module based on weight sharing is the backbone module of our proposed network, its corresponding Siamese loss will effectively optimize the parameters of the point cloud feature extraction module in the entire segmentation network. At the same time, the effectiveness improvement of smoothness loss and cross-entropy loss on the entire segmentation network is relatively weak, because the smoothness constraint has limited improvement on the robustness of semantic segmentation, and the label propagation and cross-entropy loss function only compute a small number of sampled points, resulting in their low weight. In addition, the residual connection module can help solve the problem of gradient disappearance during neural network training, and also improve the accuracy and stability of the segmentation network, which is effective in neural networks with multiple layers. Finally, the data augmentation strategy (such as different perspective transformations and adding noise) can enhance the generalization and robustness of the entire segmentation network, and improve its segmentation performance indicators.

6. Conclusions

This paper proposes a WSsegmentation network based on data augmentation and self-attention mechanism for scene point clouds. The network enhances the input scene point cloud data by performing data augmentation such as perspective transformation and adding Gaussian noise. The backbone network combines a multilayer perceptron with a self-attention mechanism to extract and aggregate data features of scene point clouds, which simultaneously combines global and local features of point clouds. Three different loss functions are introduced to optimize the network parameters during segmentation network training. Extensive experiments illustrate that the proposed segmentation network achieves good segmentation results on different point cloud datasets.

In the future, we will focus on the other strategies for enhancing point cloud data through adversarial training. Furthermore, we can investigate the semantic segmentation scheme for sparse large-scale outdoor point cloud scenes.

CRediT authorship contribution statement

Yongwei Miao: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervisor, Funding acquisition, Project administration. **Guoxiang Ren:** Methodology, Software, Investigation, Visualization. **Jinrong Wang:** Investigation, Data curation, Formal analysis, Validation. **Fuchang Liu:** Methodology, Visualization, Writing – review & editing, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant No. 61972458, and the Zhejiang Provincial Natural Science Foundation of China under Grant No. LZ23F020002.

Table 3
Comparison of mIoU evaluation by different methods on S3DIS dataset [20].

Segmentation method	Supervision mode	mIoU	Ceiling	Floor	Wall	Column	Window	Door	Chair	Table	Sofa	Board	Clutter
PointNet++ [23]	Total supervision	57.20	91.10	96.10	68.80	14.10	25.60	44.30	64.80	70.60	46.90	30.90	39.10
DGCNN [24]	Total supervision	54.27	92.70	97.45	74.60	12.90	48.20	23.50	65.20	67.20	44.20	34.30	40.30
SegCloud [6]	Total supervision	52.98	90.01	96.03	69.75	17.96	39.21	23.10	75.69	70.26	40.79	12.62	41.33
PointCNN [25]	Total supervision	60.73	92.33	98.35	79.39	15.96	22.10	63.01	81.02	74.66	31.56	63.01	56.29
k-means [26]	Unsupervised	37.45	60.00	63.20	34.70	24.30	34.00	29.40	33.90	32.90	44.90	41.30	30.80
MT [27]	Weak supervision	52.44	91.60	96.80	74.60	10.60	46.30	17.90	67.20	70.60	50.30	30.20	42.30
Xu et al. [28]	Weak supervision	52.60	90.90	97.30	74.80	8.40	49.30	27.30	69.00	71.70	53.20	23.30	42.80
Π Model [29]	Weak supervision	49.54	87.80	95.89	70.70	4.23	40.73	18.46	65.73	64.90	44.60	25.88	40.32
MulPro [30]	Weak supervision	55.64	89.70	96.90	75.50	14.00	45.70	40.70	68.50	66.80	49.40	34.40	41.20
Ours	Weak supervision	53.03	89.98	98.20	75.80	8.60	50.10	27.40	68.90	72.10	53.10	23.40	43.10

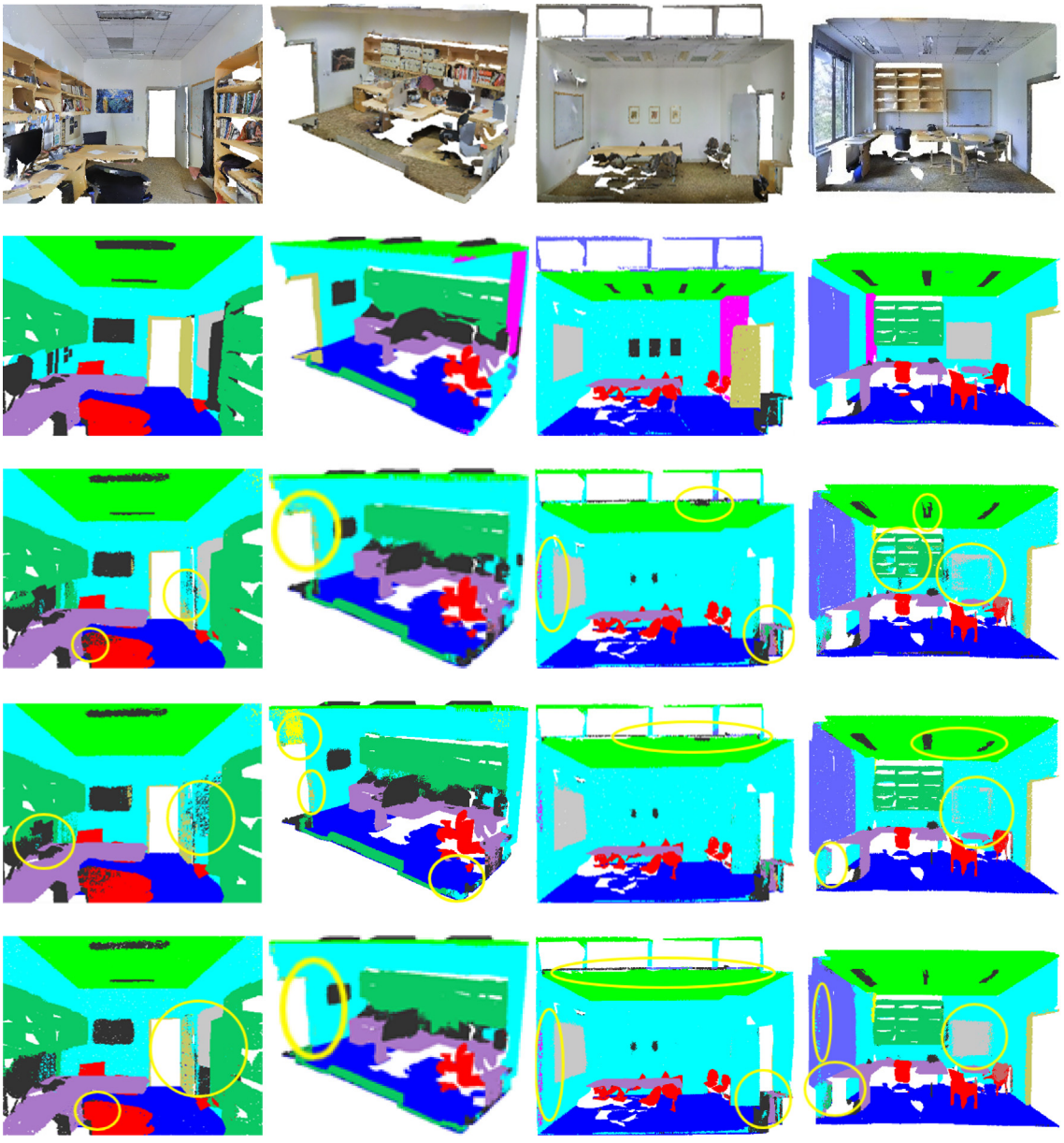


Fig. 5. Comparison of semantic segmentation results for four office scenes (a)–(d). The yellow ellipses highlight the semantic segmentation results via different methods. The first row gives the input point cloud scenes; the second row shows the ground truth of semantic labels for different point cloud scenes; the third row and the fourth row are the semantic segmentation results via PointNet++ network [23] and MT method [28] respectively; and the fifth row shows the semantic segmentation results via our proposed method.

Table 4
Ablative experiment of different components of segmentation network.

Different components						Our network	
Viewpoint transformation	Gaussian noise	Residual link	Siamese loss	Smooth loss	Cross-entropy loss	S3DIS dataset [20]	ShapeNet dataset [21]
×	×	×	×	×	×	40.16	73.23
×	×	×	×	×	√	40.55	73.53
×	×	×	×	√	√	41.37	74.01
×	×	×	√	√	√	43.97	74.76
×	×	√	√	√	√	44.27	75.10
×	√	√	√	√	√	44.53	75.51
√	√	√	√	√	√	47.32	77.42

References

[1] Miao Y, Xiao C. Geometric processing and shape modeling of 3D point-sampled models. Beijing: Science Press; 2014.

[2] Zhu Y, Mottaghi R, Kolve E, Lim JJ, Gupta A, Fei-Fei L, Farhadi A. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In: IEEE International conference on robotics and automation (ICRA), Singapore, May 29-Jun. 3. 2017, p. 3357–64.

[3] Zhang Y, Qu Y, Xie Y, Li Z, Zheng S, Li C. Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation. In: IEEE/CVF international conference on computer vision (ICCV), Montreal, Canada, Oct. 10-17. 2021, p. 15500–8.

[4] Kim S, Lee S, Hwang D, Lee J, Hwang SJ, Kim HJ. Point cloud augmentation with weighted local transformations. In: IEEE/CVF international conference on computer vision (ICCV), Montreal, Canada, Oct. 10-17. 2021, p. 528–37.

[5] Li R, Li X, Heng PA, Fu CW. PointAugment: An auto-augmentation framework for point cloud classification. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), Seattle, USA, Jun. 13-19. 2020, p. 6377–86.

[6] Tchapmi LP, Choy CB, Armeni I, Gwak J, Savarese S. Segcloud: Semantic segmentation of 3d point clouds. In: International conference on 3D vision, Qingdao, China, Oct. 10-12. 2017, p. 537–47.

[7] Gong J, Ye Z, Ma L. Neighborhood co-occurrence modeling in 3d point cloud segmentation. Comput Vis Media 2022;8(2):303–15.

[8] Zhao Y, Ma X, Hu B, Zhang Q, Ye M, Zhou G. A large-scale point cloud semantic segmentation network via local dual features and global correlations. Comput Graph 2023;111:133–44.

[9] Su H, Maji S, Kalogerakis E, Learned-Miller E. Multi-view convolutional neural networks for 3d shape recognition. In: IEEE international conference on computer vision (ICCV), Santiago, Chile, Dec.7-13. 2015, p. 945–53.

[10] Hou J, Dai A, Nießner M. 3D-SIS: 3D semantic instance segmentation of RGB-D scans. In: IEEE conference on computer vision and pattern recognition (CVPR), Long Beach, USA, Jun. 16-20. 2019, p. 4421–30.

[11] Nan L, Xie K, Sharf A. A search-classify approach for cluttered indoor scene understanding. ACM Trans Graph 2012;31(6):137:1–137:10.

[12] Li Y, Dai A, Guibas L, Nießner M. Database-assisted object retrieval for real-time 3d reconstruction. Comput Graph Forum 2015;34(2):435–46.

[13] Mattausch O, Panozzo D, Mura C, Sorkine-Hornung O, Pajarola R. Object detection and classification from large-scale cluttered indoor scans. Comput Graph Forum 2014;33(2):11–21.

[14] Wu Y, Cai S, Yan Z, Li G, Yu Y, Han X, Cui S. Pointmatch: A consistency training framework for weakly supervised semantic segmentation of 3d point clouds. 2022, arXiv:2202.10705, 2022.

[15] Guo MH, Cai JX, Liu ZN, Mu TJ, Martin RR, Hu SM. PCT: Point cloud transformer. Comput Vis Media 2021;7(2):187–99.

[16] Vanian V, Zamanakos G, Pratikakis I. Improving performance of deep learning models for 3d point cloud semantic segmentation via attention mechanisms. Comput Graph 2022;106:277–87.

[17] Hu Q, Yang B, Fang G, Guo Y, Leonardis A, Trigoni N, Markham A. Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds.. In: European conference on computer vision (ECCV), Tel Aviv, Israel; Oct. 23-27. 2022, p. 600–19.

[18] Zhou Z, Zhou Y, Wang D, Mu J, Zhou H. Self-attention feature fusion network for semantic segmentation. Neurocomputing 2021;453:50–9.

[19] Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), San Diego, USA, Jun. 20-25. 2005, p. 539–46.

[20] Armeni I, Sener O, Zamir AR, Jiang H, Brilakis I, Fischer M, Savarese S. 3D semantic parsing of large-scale indoor spaces. In: IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, USA, Jun. 27-30. 2016, p. 1534–43.

[21] Chang AX, Funkhouser TA, Guibas LJ, Hanrahan P, Huang Q, Li Z, Savarese S, Savva M, Song S, Su H, Xiao J, Yi L, Yu F. ShapeNet: An information-rich 3d model repository. 2015, http://dx.doi.org/10.48550/arXiv.1512.03012, arXiv:1512.03012, 2015.

[22] Mo K, Zhu S, Chang AX, Yi L, Tripathi S, Guibas LJ, Su H. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In: IEEE conference on computer vision and pattern recognition (CVPR), Long Beach, USA, Jun. 16-20. 2019, p. 909–18.

[23] Qi CR, Yi L, Su H, Guibas LJ. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: Annual conference on neural information processing systems (NeurIPS), Long Beach, USA, Dec. 4-9. 2017, p. 5099–108.

[24] Phan AV, Nguyen ML, Nguyen YLH, Bui LT. DGCNN: A convolutional neural network over large-scale labeled graphs. Neural Netw 2018;108:533–43.

[25] Li Y, Bu R, Sun M, Wu W, Di X, Chen B. Pointcnn: Convolution on X-transformed points. In: Annual conference on neural information processing systems (NeurIPS), Montréal, Canada, Dec. 3-8. 2018, p. 828–38.

[26] Dhanachandra N, Manglem K, Chanu YJ. Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. Procedia Comput Sci 2015;54:764–71.

[27] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Annual conference on neural information processing systems (NeurIPS), Long Beach, USA, Dec. 4-9. 2017, p. 1195–204.

[28] Xu X, Lee GH. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), Jun. 16-18. 2020, p. 13706–15.

[29] Laine S, Aila T. Temporal ensembling for semi-supervised learning. 2016, http://dx.doi.org/10.48550/arXiv.1610.02242, arXiv:1610.02242, 2016.

[30] Su Y, Xu X, Jia K. Weakly supervised 3D point cloud segmentation via multi-prototype learning. IEEE Trans Circuits Syst Video Technol 2023. http://dx.doi.org/10.1109/TCSVT.2023.3281151.