Article

Weakly Supervised Semantic Segmentation of Point Cloud Scenes using Boundary-based Feature Aggregation

Yongwei Miao , Guoxiang Ren , Xudong Zhang * , Haijian Liu , Fuchang Liu *

Posted Date: 1 April 2024

doi: 10.20944/preprints202404.0038.v1

Keywords: Point cloud; Semantic segmentation; Weakly supervised learning; Boundary feature aggregation; Data augmentation

Article

Weakly Supervised Semantic Segmentation of Point Cloud Scenes Using Boundary-Based Feature Aggregation

Yongwei Miao [1], Guoxiang Ren [1], Xudong Zhang [2,*], Haijian Liu [1] and Fuchang Liu [1,*]

[1] School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China
[2] School of Information Science and Technology, Zhejiang Shuren University, Hangzhou 310015, China
* Correspondence: xdzhang@zjsru.edu.cn; liufc@hznu.edu.cn

Abstract: The task of weakly supervised point cloud semantic segmentation has received widespread attention and also has been widely used in autonomous driving, robotics, and modern industries. Due to the high dimensionality and large volume of point cloud data, many technical difficulties appeal to weakly supervised semantic processing, especially segmentation. However, weakly-supervised schemes usually provide only partial labelling information of the underlying point cloud data and thus need to effectively extract local features, geometric information, and contextual relationships only using these limited labelled data for supervised learning. To improve weakly-supervised semantic segmentation, we propose a novel segmentation network through the boundary-based feature aggregation based on a K-NN algorithm with down-sampling operation, and also introduce the smoothness loss and Siamese loss for effective segmentation. The experiments on public datasets demonstrate that our presented segmentation method exceeds most of the existing fully supervised and weakly-supervised methods in terms of mIoU. Specifically, our network has high segmentation accuracy on the labels of objects with similar geometrical structures, such as ceiling, wall, floor, chair and table, reaching 91.2%, 98.8%, 83.3%, 75.3% and 80.2%, respectively. Extensive experiments also illustrate its robustness, effectiveness, and generalization of the proposed weakly supervised segmentation network.

Keywords: point cloud; semantic segmentation; weakly supervised learning; boundary feature aggregation; data augmentation

1. Introduction

Weakly supervised semantic segmentation of point cloud scenes is a significant area of research within the fields of computer graphics and 3D computer vision. This is primarily due to the time-intensive and laborious properties of annotating point clouds. Existing methodologies for point cloud semantic segmentation in traditional contexts often encounter challenges when processing scene boundaries across diverse semantic categories [1–3]. A particularly demanding aspect involves the segmentation of scene boundaries, which requires distinguishing different semantic labels by sampling point features based on approximate color and geometric information. This process is crucial for the accurate segmentation of various objects within indoor point cloud scenes.

Given the expansive nature of large-scale scene datasets, segmentation approaches increasingly rely on weakly-supervised learning strategies [4,5]. These strategies are essential for efficiently segmenting extensive point cloud data and identifying diverse object labels within complex scenes, particularly when semantic annotation information is limited. Although current weakly supervised methods for point cloud semantic segmentation yield effective results in scene segmentation, they often struggle to accurately segment regions near scene boundaries. For the applications such as autonomous driving, the precise detection of objects and road edges is imperative for safety and efficiency [6,7]. However, current 3D segmentation methodologies often tend to ignore the feature extraction of scene boundaries. This aspect is of critical importance in these applications, as it directly impacts the system's ability to accurately interpret and navigate its surroundings.

Furthermore, the manual labeling of extensive point cloud data is not only time-consuming but also incurs significant costs. Weakly supervised methods offer a viable solution by reducing the dependency of neural networks on the quantity of point cloud labels. This reduction in label requirements subsequently lowers both the cost and complexity of point cloud segmentation, facilitating its practical application and wider adoption.

Addressing the aforementioned challenges, this paper presents a novel approach involving a boundary-based sampling, feature extraction, and aggregation module. This module is seamlessly integrated into a weakly-supervised network for point cloud semantic segmentation, with a special emphasis on enhancing boundary segmentation. This innovation significantly elevates the overall segmentation precision of point clouds and robustly fortifies boundary segmentation capabilities. Beyond point sampling and data augmentation at boundaries, our methodology further incorporates Gaussian noise data augmentation and rotation sensitivity enhancement. These strategic augmentations remarkably strengthen the model's robustness and accuracy across a spectrum of scenarios.

The main technical contributions of this work are summarized as follows:

- An innovative module specifically tailored for efficient boundary sampling method is proposed, which leverages a K-NN algorithm combined with multiscale downsampling operations, significantly enhancing the precision of boundary delineation and facilitating improved aggregation of scene boundary features.
- An advanced feature extraction and aggregation method is incorporated into the weakly-supervised point cloud segmentation network. This method is based on linear self-attention modules, which is characterized by fewer parameters and rapid inference speeds and improves the segmentation quality.
- We have also conducted a comprehensive study on data augmentation for weakly annotated point clouds. This study includes the application of Gaussian noise data augmentation and the enhancement of rotation sensitivity. These strategic augmentations significantly contribute to the robustness and adaptability of the model, ensuring its effective performance under diverse application scenarios.

2. Related Work

Weakly supervised learning strategies have emerged as an effective solution to address the challenges of limited labeling information and to reduce the associated costs of labeling. This approach has gained considerable attention in the realm of point cloud scene understanding [4,8,9]. In this paper, we present a comprehensive review of the related work focusing on point cloud data enhancement and semantic segmentation, particularly emphasizing developments in weakly supervised learning.

2.1. Data Augmentation of Point Cloud

In the context of weakly supervised learning, point cloud data enhancement typically encompasses techniques such as random rotation, model scaling, and color dithering [10]. These enhancements significantly improve the generalization capabilities of neural network training. For instance, random rotation involves the 3D rotation of point cloud data, with random selection of the axis and rotation. This process generates varied model perspectives and poses, enhancing the model's ability to identify and classify targets and increase network robustness to poses changing. Model scaling, on the other hand, entails the 3D scaling of point cloud data. By randomly varying the scaling factor, the model gains robustness to changes in target sizes. Color dithering involves transformations in the color attributes of point cloud data, including adjustments in luminance, contrast, and saturation, thereby increasing data diversity and improving the model's robustness in color differentiation. Kim et al. [11] employed a local weighting transform to produce non-rigid deformations in point clouds. This kernel density-based estimation method assigns weights to each sample point inversely proportional to their distance from target points, enhancing the accuracy of the transformed region while preserving the global shape of the point cloud. Li et al. [12], using an adversarial learning approach, proposed an automatic enhancement framework for hyperspectral image data. This framework utilizes Gaussian noise and masking techniques to generate enhanced hyperspectral maps, filtering noise through randomly generated binary masks to accentuate image features. Leng et al. [13] introduced three pseudo-label-based data augmentation policies (PseudoAugments) to integrate both labeled and pseudo-labeled scenes. Their approach leverages unlabeled data for data augmentation, enriching the training dataset for the network. Wang et al. [14] developed a rotation-based point cloud adversarial attack method, which points out the sensitivity of many point cloud networks to the rotation of point clouds. This observation has inspired further research into the impact of point cloud rotation augmentation on weakly supervised semantic segmentation networks. Miao et al. [15] developed a deep learning-based weakly supervised semantic segmentation network for indoor scene point clouds. Utilizing three loss functions and data augmentation with a minimal label set, their methodology achieves a network structure characterized by enhanced segmentation efficiency and reduced time consumption. This is achieved through the simplification of point cloud boundary features, narrowing the rotation angle range during data augmentation, and optimizing the loss function.

However, the above method mainly focuses on discrete point cloud data, it is difficult to simulate the effects of different light intensities, rotation robustness and point cloud noise on point cloud data obtained in real applications, which makes it difficult for the neural network to achieve ideal semantic segmentation results during training. In this paper, the point cloud semantic segmentation network is augmented with rotation and Gaussian noise to generate diverse point cloud samples for training indoor objects.

2.2. *Weakly Supervised Semantic Segmentation of Point Cloud Data*

Neural network-based segmentation strategies have shown significant advantages in various semantic segmentation tasks [16,17]. Leveraging the success of convolutional neural networks in feature extraction on images, most approaches typically employ view-based [17] or volume representation [16] techniques to convert 3D discrete point cloud data into regular mesh data. For instance, Zhou et al.[17] transformed sampling points within each voxel into feature vectors for effective extraction, applying this to target detection tasks. Owing to the multi-modal collaborative learning, Ni et al. [18] proposed a robust 3d semantic segmentation method thus to enhance feature extraction and segmentation performance for point clouds. Despite their effectiveness, multi-view image data representations have limitations due to object occlusion and complex data inputs, while voxel representations are often limited by resolution constraints due to the high storage demands.

Indoor scene point cloud semantic segmentation methods primarily fall into model-based and primitive-based strategies. Nan et al. [19] introduced a scene modeling method using pre-trained object categories for search classification and segmentation. Li et al. [20] proposed replacing scanned 3D data with objects from a shape database for scene understanding and reconstruction. Shi et al. [21] segmented indoor scenes by training classifiers for object and group classification. While these methods effectively utilize database shape information, they often rely heavily on the diversity and size of the model dataset. An alternative approach involves decomposing indoor scenes into geometric primitives. Mattausch et al. [22] detected duplicate objects in scenes using numerous planar slices, clustering them with a geometric similarity matrix for scene segmentation. Wu et al. [23] proposed a consistency training framework, generating pseudo-labels from perturbed and rotated unlabeled point clouds to enhance model robustness and generalization, thus preventing overfitting. Hu et al. [24] developed a weakly-supervised semantic segmentation method for large-scale discrete point clouds, generating virtual labeled data from known shape and color information to train a classifier for semantic segmentation.

The proposed weakly-supervised point cloud semantic segmentation network in this paper improves the generalization and robustness of traditional methods based on data enhancement strategies, offering more accurate semantic segmentation results than the existing end-to-end methods. Our network architecture eliminates the necessity of selecting clustering algorithms and initializing cluster numbers, thereby also circumventing the need for regularization. The neural network weights

doi:10.20944/preprints202404.0038.v1

are continuously optimized through a multi-layer perceptron and self-attention mechanism, which ensures the effectiveness and efficiency of point cloud segmentation.

3. Weakly Supervised Point Cloud Semantic Segmentation Network

3.1. Overview of Our Method

The existing weakly supervised point cloud semantic segmentation networks often struggle with challenges such as handling boundary sampling point labels, sensitivity to label noise, and limited generalization of point cloud data, leading to suboptimal accuracy in point cloud semantic segmentation. To address these issues, this paper proposes a novel weakly supervised point cloud scene semantic segmentation network, which emphasizes boundary processing. The overarching framework of this network is illustrated in Figure 1. Initially, the network randomly rotates the original point cloud data within a restricted angle range and introduces Gaussian noise without parameter value limitations, creating a dataset with enhanced feature contrast. Subsequently, utilizing the K-NN algorithm, points surrounding the sampling points are selected for downsampling, yielding point clouds that predominantly reflect the classes of the original sampling points. Further, through self-supervision, the point cloud data undergoes an additional downsampling step to retain the original color characteristics of the sampling points. The network then leverages a self-attention model and a multi-layer perceptron as part of the point cloud feature extraction module, effectively extracting and aggregating features from the enhanced point clouds. Finally, the network employs a maximum pooling layer and a multi-layer perceptron to downscale the point cloud features, culminating in the generation of the point cloud segmentation results.

In the training stage, the neural network incorporates the Siamese loss function and the Smooth loss function to enhance the accuracy of point cloud semantic segmentation. The Siamese loss function is utilized to train twin neural networks on point cloud labels, thereby significantly improving the segmentation precision of point cloud scenes. Concurrently, the Smooth loss function is employed to penalize the label discontinuity among neighboring samples that share similar geometric and color information. This approach is instrumental in further refining the segmentation accuracy of point cloud scenes.

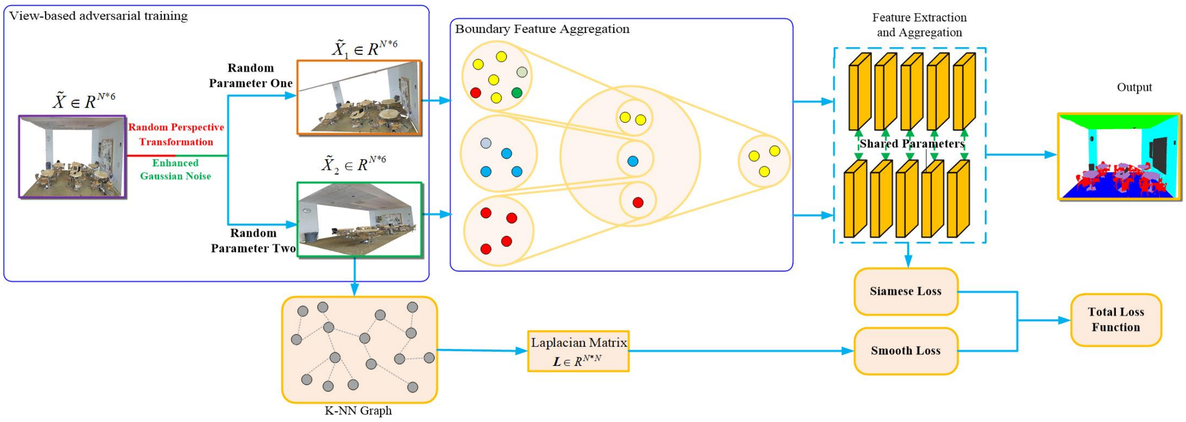

Figure 1. The proposed weakly supervised semantic segmentation network. The input point cloud is first fed to data augmentation through Gaussian noise and rotation-aware techniques. Subsequent to this augmentation, boundary feature extraction and aggregation is performed, ultimately yielding the semantic segmentation results. Different colors of small circles in the middle of the figure represent point clouds of various categories.

3.2. Data Augmentation and Sampling

To significantly enhance the effectiveness and accuracy of point cloud scene semantic segmentation, the size of the scene dataset is a crucial factor in neural network model training. A limited number of real scene point cloud data can substantially restrict the model's generalization ability. Therefore, effective enhancement of point cloud scene data is essential [10,11,25,26]. Augmentation of scene point cloud data can be achieved through various transformations applied to the original point cloud scenes, such as rotational transformation [11] and the addition of Gaussian noise [24]. They are employed to generate new scene point cloud datasets for increasing the diversity of the dataset, which is key to ensuring the generalization capability of the segmentation network. By enhancing point cloud data to create a range of differently transformed scene data, the segmentation model is improved to identify and distinguish between various classes of point cloud objects. Consequently, this leads to improved generalization abilities of the network.

3.2.1. Gaussian Noise Enhancement

In the realm of weakly-supervised point cloud semantic segmentation, the scarcity of scene point cloud data samples, coupled with the high costs of annotating such data, often results in insufficient training samples for neural network training. This limitation frequently leads to model overfitting, where the network performs well on training data but falls short on test data. To mitigate this issue, our research further explores the enhancement of point cloud data. This enhancement involves adding Gaussian noise to the rotationally transformed point cloud data, thereby enriching the scene sample data. In the context of point cloud scene segmentation tasks, Gaussian noise effectively simulates inaccuracies and semantic segmentation noise inherent in scene data, which might arise from the sampling equipment and methods used during data acquisition. By incorporating Gaussian noise into scene point cloud data, the trained network model becomes more robust to data noise, enhancing the diversity of the scene dataset and consequently improving the model's generalization performance. Specifically, for scene point cloud data represented as $\boldsymbol{P} = (x_i, y_i, z_i), i = 1, 2, \cdots, n$, each sampling point's coordinates are perturbed with random fluctuations conforming to a normal distribution, as follows.

$$\begin{aligned} x_i &= x_i + \sigma_x \cdot \xi \\ y_i &= y_i + \sigma_y \cdot \xi \\ z_i &= z_i + \sigma_z \cdot \xi \end{aligned} \tag{1}$$

Where, the variable ξ obeys the standard normal distribution $N(0,1)$, σ_x, σ_y, σ_z are the standard deviations of the added noise respectively.

In the process of point cloud data enhancement, the combination of perspective transformation and the addition of Gaussian noise simultaneously improves the efficacy of data enhancement. Specifically, the original point cloud undergoes a transformation, followed by an enhancement through the addition of Gaussian noise. This approach ensures the presence of noise across various angles, thereby substantially increasing the data's diversity.

3.2.2. Rotation-Aware Data Enhancement

The primary objective of a point cloud semantic segmentation model is to discern and extract pivotal feature information from point clouds, enabling the accurate classification of each sample point within the point cloud scene. However, the inherent irregularity of point cloud data poses a significant challenge to effective feature extraction, subsequently impacting the accuracy of neural network segmentation. To address this challenge, it is imperative for the point cloud semantic segmentation network to perform data enhancement on the original point cloud prior to network training. A practical enhancement strategy involves the random rotation of field point cloud data called rotation-aware data enhancement. Implementing this method is crucial as it facilitates more efficient feature extraction from point cloud data, thereby substantially boosting the generalization performance of the segmentation model, especially for point clouds with irregular distributions.

To maximize the generalizability of this data enhancement method on the sampled points, we have configured the rotation matrix $\boldsymbol{R}$ to conform to a Bernoulli distribution $B(1, 0.5)$. Simultaneously, we ensure that the rotation angle θ adheres to a uniform distribution $U(0, 0.5\pi)$. This configuration guarantees that the rotation angles are uniformly distributed within the specified range. For instance,

$$R_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & \sin\theta_x & \cos\theta_x \end{bmatrix} \tag{2}$$

$$R_y = \begin{bmatrix} \cos\theta_y & 0 & \sin\theta_y \\ 0 & 1 & 0 \\ -\sin\theta_y & 0 & \cos\theta_y \end{bmatrix} \tag{3}$$

$$R_z = \begin{bmatrix} \cos\theta_z & -\sin\theta_z & 0 \\ \sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

$$R = R_z(\theta_z) \cdot R_y(\theta_y) \cdot R_x(\theta_x) \tag{5}$$

Here, the first transformation matrix in our approach governs the rotation degree of the original point cloud data, while the second matrix controls the magnitude of the coordinate transformation between the point cloud image and the original data. Enhanced point cloud data samples are derived by multiplying the original point cloud data $\boldsymbol{X}$ with the transpose of the rotation matrix, expressed as $\widehat{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{R}^T$. This rotation transformation pivots the scene's point cloud data across various viewing directions, thereby generating data from multiple perspectives. Such an approach aids the segmentation model in more effectively identifying and distinguishing different objects within the scene. Furthermore, it contributes to an increase in the quantity of scene point cloud data samples, thereby expanding the pool of sample data available for model training.

3.2.3. Boundary-Based Sampling

The K-NN (K-Nearest Neighbors) algorithm is a widely utilized method in point cloud data processing, predominantly employed for the selection of sampling points and the identification of the nearest K points based on the shortest Euclidean distance. This algorithm plays a pivotal role in determining the local structure and boundaries within a point cloud scene. This objective is accomplished by evaluating the similarity in color and geometric information between the identified sample points and their adjacent points.

In practice, the neural network employs uniform sampling in conjunction with the K-NN algorithm to ascertain the sampling points and their nearest K neighbors. The process involves a comparative analysis of the feature information of several sampled points against their contextual relationships to discern whether they belong to the boundaries of indoor scenes or are objects within the point cloud. The K-NN algorithm typically utilizes Euclidean distance or other relevant distance metrics for its calculations. The formulation of the algorithm is as follows:

$$NN(i) = argmin_j \|\boldsymbol{P}_i - \boldsymbol{P}_j\| \tag{6}$$

Following the selection of sampling points via the K-NN algorithm, our network employs downsampling operations to decrease the number of points, both in terms of the sampling points and their surrounding neighbors. This reduction in point quantity effectively minimizes computational complexity and enhances processing efficiency. Commonly employed downsampling techniques encompass voxel grid downsampling, random sampling, and distance-based sampling. As illustrated in Figure 2, the neural network discerns the point cloud boundary using geometric and color information. It then determines whether different sampling points correspond to the same point cloud label by assessing the magnitude of the Euclidean distance in their color information. This process simplifies the complex color information present at the point cloud boundaries via downsampling, thereby facilitating a more straightforward segmentation task.

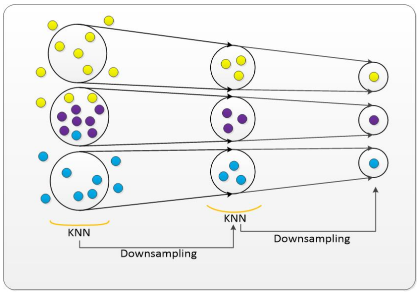

Figure 2. Boundary feature difference aggregation module

3.3. *Self-Attention Based Feature Extraction and Aggregation*

Typically, to augment the feature representation capabilities of point cloud data, neural network models often incorporate self-attention mechanisms and multilayer perceptrons for feature extraction [27]. The self-attention mechanism, as an efficacious form of point cloud feature characterization, focuses on distinct feature information from various sampling points in the scene. Concurrently, it learns the intrinsic local relationships between each sampling point and its counterparts, thereby more effectively capturing the global contextual information within the scene's point cloud data. Additionally, the structural parameters of this neural network are fewer compared to conventional networks, which translates to heightened efficiency in extracting complex local features from point clouds. The multilayer perceptron, functioning as a type of nonlinear transformation, maps point cloud features into a higher-dimensional feature space. This mapping is crucial for achieving an effective representation of high-dimensional features. Moreover, the layering of the perceptron enables the neural network to progressively abstract higher-level feature information from the point cloud, thereby enhancing the model's representational capacity. We have effectively integrated the self-attention mechanism with the multilayer perceptron. This integration optimizes the capture of both global and local features of the site point cloud data, significantly improving the accuracy and robustness of the semantic segmentation of the site point cloud. Moreover, this approach significantly enhances the model's discriminative power and its capacity for generalization.

Here, we introduce a feature extraction and aggregation structure leveraging the self-attention mechanism, adeptly combining local and global features of point clouds. Initially, the enhanced point cloud data, configured as an $N * 6$ dimensional input where N represents the number of sampling points, is processed. Each point in this dataset is characterized by a 6-dimensional vector (x, y, z, r, g, b), encapsulating its spatial coordinates and color information. The multi-layer perceptron (MLP) first maps each point's low-dimensional features into a 64-dimensional feature space, enabling the network to more effectively discern both local structures and global features of the cloud. Subsequently, these high-dimensional features are introduced into a Self-attention (SA) model, which along with the MLP, extracts and aggregates the point cloud features. The 64-dimensional features are then processed through the Self-attention module. This module adaptively calculates the weights between sampling points based on their features, aggregating this information. Its primary goal is to bolster the interconnectedness and interaction among point cloud features, thereby enhancing the accuracy and robustness of their representation. Furthermore, the use of self-attention models significantly trims the count of neural network parameters compared to using only MLPs. For example, an MLP with an input dimension of 64, a hidden layer of 128, and an output dimension of 512 requires 74 KB parameters. In contrast, a self-attention structure with similar dimensions and three attention heads only necessitates 36 KB parameters, nearly halving the computational load.

This mechanism facilitates the interaction among feature vectors from each point within the scene, thereby effectively capturing the global information of the point cloud. The self-attention module, in comparison to traditional methods, demonstrates enhanced proficiency in extracting both local and global structures of the cloud. This attribute renders it highly effective for interpreting a diverse range of point cloud scenes. In the fusion of local and global features, our study strategically concatenates various scales of point cloud information along the channel dimension prior to feeding it into the neural network for semantic segmentation. During the neural network training process, challenges such as gradient explosion or gradient vanishing often pose difficulties in training the model effectively. As depicted in Figure 3, the network structure employs a multi-layer perceptron to transition enhanced point cloud data from a 6-dimensional space to a 64-dimensional space. The point cloud features within this dimension are then processed through an attention model. After extracting local feature data from complex point clouds, we obtain new 64-dimensional point cloud features. A repetition of this process further enhances these features to a 1024-dimensional space, thereby enhancing the local characteristics.

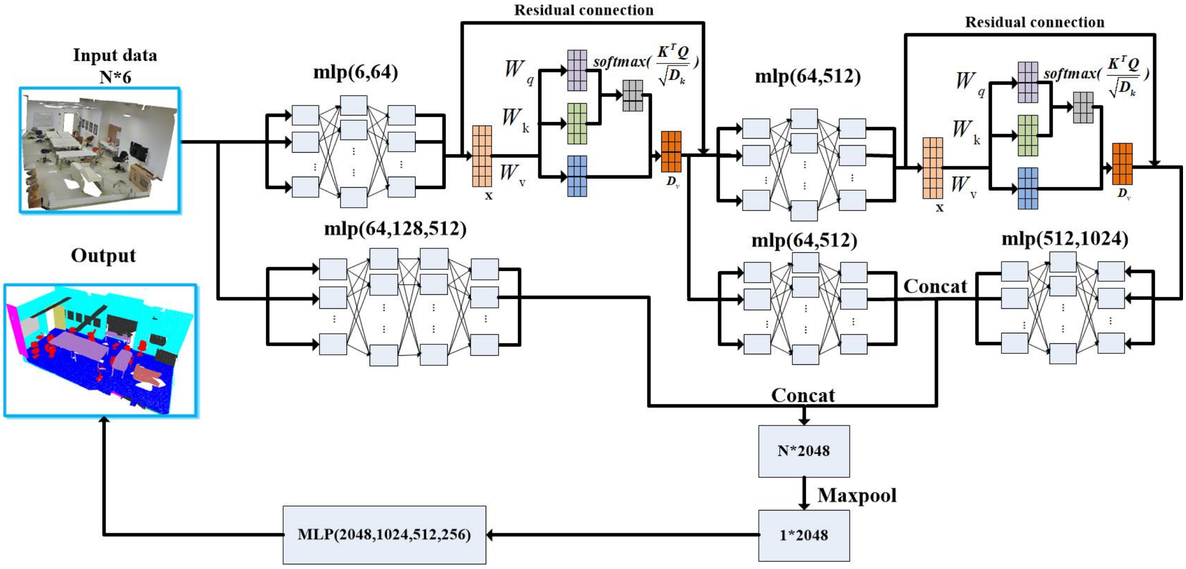

Figure 3. Feature extraction and aggregation module. Three multi-layer perceptrons are utilized to conduct dimensionality enhancement operations on the input point cloud data. These operations are integrated with the point cloud features extracted from two successive rounds of local feature extraction. As a result, this process yields two sets of 512-dimensional features and one set of 1024-dimensional point cloud features.

4. Experimental Results and Discussions

The weakly supervised point cloud semantic segmentation network presented in this paper was implemented on a system running Ubuntu 16.04, equipped with an Intel Core i9-10900K CPU processor at 3.70GHz, 500GB RAM, and an NVIDIA GTX 1080 GPU. The software development was conducted using Python 3.6 and PyTorch 1.6.

For training, the network utilized datasets of S3DIS [10] and ShapeNet [25] with only 1.0% data labels. This study primarily focuses on the accuracy of point cloud boundary segmentation within the segmentation results. To enhance the generalization of a limited set of point cloud labels, we conducted experiments involving two data enhancement operations. Additionally, the boundary feature disparity aggregation module was employed to simplify the segmentation of boundary samples. The network leverages a local self-attention mechanism and a multi-layer perceptron to capture the local and global features of scene scanning data from point clouds. High-dimensional point cloud features are then fused and processed through maximal pooling operations to predict unlabelled sample points in the point cloud scene.

During the network training phase, the parameters of our segmentation network were optimized using a combined loss function of Smooth and Siamese losses to enhance the accuracy of semantic segmentation. In the testing phase, the network followed the same procedural steps as in training for category inference, culminating in the final semantic segmentation results for the scene point cloud data.

4.1. Evaluation of the Proposed Network

In this study, we have trained and tested a weakly supervised point cloud semantic segmentation network, which incorporates a boundary feature difference aggregation module, on the S3DIS dataset Area 5 [10]. The testing duration for this area was approximately 3.5 hours, and the neural network's parameter memory footprint was around 80MB. This dataset presents significant challenges in various scenarios.

For instance, as depicted in the first row of Figure 4, elements such as the floor, indoor chairs and table can be effectively segmented, whilst the whiteboard on the wall also can be segmented though it has similar geometry and colour information with the wall. In the second row of Figure 4, the ceiling, floor, chairs and tables can be effectively segmented. In the third row of Figure 4, elements for example the ceiling, floor, table, the overhead light, the bookcases on both sides and the picture on the wall are effectively segmented. Lastly, in the final row of Figure 4, the ceiling, floor, table, the bookcases on both sides and indoor chairs are effectively segmented. The K-NN algorithm introduced in our segmentation network, enhanced with a downsampling operation, demonstrates improved accuracy in weakly supervised point cloud semantic segmentation for indoor scenes.

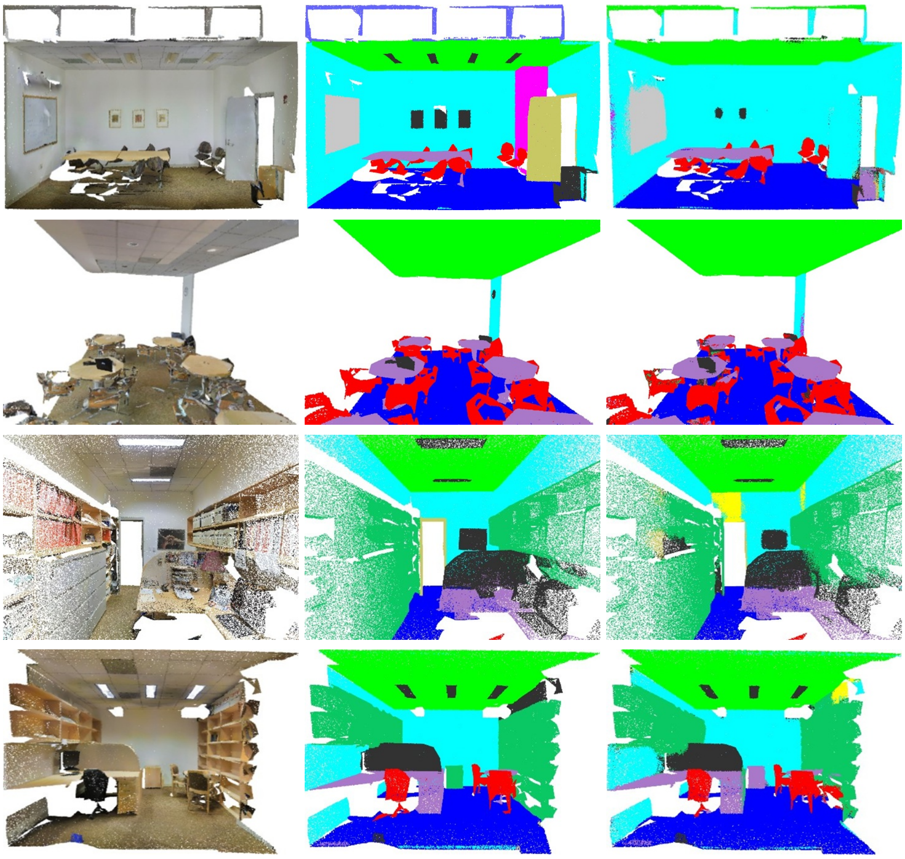

Figure 4. The segmentation results of our proposed weakly supervised semantic segmentation network are presented as follows: The first column displays the input model, the second column illustrates the corresponding ground truth of semantic labels, and the third column depicts the segmentation results achieved through our proposed network.

4.2. Comparison with State of the Art

In this experiment, thirteen methods including but not limited to weakly supervised point cloud semantic segmentation (PointNet++ [28], SegCloud [27], PointCNN [29], SPGraph [30], PCT [31], HPEIN [32], MinkowskiNet [33], KPConv [34], JSENet [35], CGA-Net [36], RandLA-Net [37], CloserLook3D [38], and Ours method) to compare the effect of semantic segmentation on the S3DIS dataset Area 5 [10], and the results are shown in Table 1.

Table 1. Comparison of segmentation results between our method and related methods on the S3DIS Area 5 dataset [10]

Method	Supervision	**mIoU**	ceiling	floor	wall	column	window	door	table	chair
PointNet++ [28]	Full	52.9	87.9	96.3	69.1	3.8	46.3	9.9	58.1	51.6
SegCloud [27]	Full	60.3	89.8	96.2	69.7	19.2	39.1	22.9	70.1	75.6
PointCNN [29]	Full	65.5	92.1	97.9	78.3	17.1	22.6	61.1	73.9	79.9
HPEIN [32]	Full	69.9	90.8	97.9	80.9	22.9	64.8	39.8	74.8	86.9
MinkowskiNet [33]	Full	73.7	91.8	97.7	86.3	33.9	47.9	61.9	80.9	88.8
KPConv [34]	Full	74.1	92.7	96.9	81.9	22.8	57.9	68.8	80.9	90.8
JSENet [35]	Full	74.6	93.7	96.9	83.1	22.9	60.8	70.8	88.8	78.8
CGA-Net [36]	Full	75.0	93.1	98.1	82.9	24.9	58.8	70.1	89.8	81.9
RandLA-Net [37]	Full	69.8	90.9	94.8	80.1	23.9	61.9	47.0	76.1	82.9
CloserLook3D [38]	Full	74.0	93.9	98.7	81.9	24.8	50.8	70.1	91.1	80.7
SPGraph [30]	Weak	71.6	89.1	95.9	77.9	41.9	47.9	61.8	83.9	74.3
PCT [31]	Weak	69.7	89.98	98.1	79.9	18.7	61.0	47.9	75.9	84.9
Ours	Weak	69.9	91.2	98.8	83.3	25.9	58.2	46.3	75.3	80.2

The experimental results use the average intersection and merger ratio (mIoU[35]) as a performance metric to evaluate and compare the performance metrics of thirteen methods on eight different categories. The results show that the semantic segmentation of this experiment is higher than RandLA-Net [37], PCT [31], PointNet++ [28], SegCloud [27], PointCNN [29], and demonstrates the same point cloud semantic segmentation accuracy as HPEI [32] on mIoU [39] metrics, and at the same time, the segmentation accuracy of this method is lower than SPGraph [30], MinkowskiNet [33], KPConv [34], JSENet [35], CGA-Net [36], CloserLook3D [38]. Specifically, the neural network in this paper has high segmentation accuracy on the labels of objects with similar geometrical structures, such as ceiling, wall, floor, chair and table, while it can obtain higher segmentation metrics on point cloud data with significant differences in colour information, reaching 91.2%, 98.8%, 83.3%, 75.3% and 80.2%, respectively.

Figure 5 shows the comparison between the SegCloud network [27], PointCNN [29] and our proposed network for semantic segmentation of indoor point cloud scenes in Area 5 of the S3DIS dataset [10], respectively. It can be seen that the SegCloud network [27] has higher segmentation accuracy than the traditional weakly-supervised point cloud semantic segmentation network, but it is still difficult to obtain the semantic labels of the overhead lights in the third and fourth rows. The PointCNN [29] network has higher point cloud segmentation indexes than the SegCloud network [27], but it still has some shortcomings in the wall segmentation in the first, second and fourth rows. The method in this paper not only has higher point cloud segmentation index than the previous two methods, but also the graphical effect of the point cloud segmentation results of this method is better than the previous two methods.

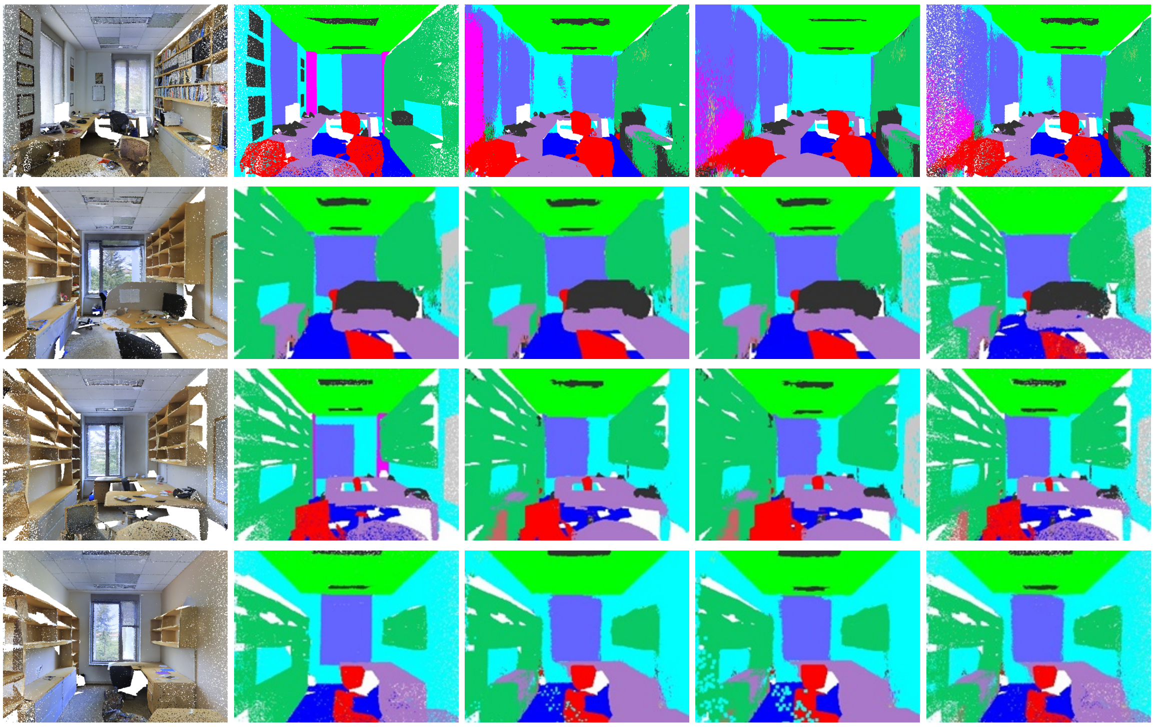

Figure 5. Comparative analysis of various point cloud semantic segmentation methods is presented for each point cloud scene as follows: The first column displays the input model, and the second column illustrates the corresponding ground truth of semantic labels. The third column depicts the segmentation results obtained using SegCloud [27]. The fourth column demonstrates the segmentation results achieved with PointCNN [29]. Finally, the last column showcases the segmentation results via our proposed network.

4.3. Ablation Study

To validate the effectiveness of the various components proposed in this paper, including the data enhancement strategy, residual linking module, Siamese loss, smoothness loss, and cross-entropy loss, we conducted ablation experiments. These experiments specifically focus on the K-NN algorithm with downsampling operation, data enhancement, residual linking, and the two loss functions. The objective is to ascertain the contribution and impact of each module on point cloud semantic segmentation. The results are presented in Table 2 by using mIoU [39] metrics on S3DIS [10] and ShapeNet datasets [25] in order to demonstrate the effect of different network modules on the accuracy of semantic segmentation results for point clouds respectively.

From Su et al. [39], it can be observed that the Boundary Feature Difference Aggregation (BFDA) module improves the mIoU metrics by 2.2 and 4.21 for the S3DIS [10] and ShapeNet datasets [25], respectively.In addition to this, the point cloud rotation and Gaussian noise have the greatest impact on the neural network's point cloud segmentation accuracy, which increases the mIoU metrics by 4.6 and 4.0, respectively. Residual linking has the smallest effect. Siamese loss and Smooth loss on the training of the neural network can significantly improve the point cloud segmentation accuracy of this neural network on both datasets.

Table 2. Ablative experiment for different components of segmentation network

Different components					Our network	
Data augmentation	BFDA module	Residual Link	Twinning Loss	Smooth Loss	S3DIS Dataset [10]	ShapeNet Dataset[25]
×	×	×	×	×	57.3	44.5
×	×	×	×	√	58.3	47.1
×	×	×	√	√	62.1	49.8
×	×	√	√	√	63.1	50.4
×	√	√	√	√	65.3	54.6
√	√	√	√	√	69.9	58.6

5. Conclusion

In this study, we have developed a weakly supervised point cloud semantic segmentation network, incorporating a boundary feature sampling and aggregation module. Within this network, the K-NN algorithm is employed to identify sampling points and their nearest neighbors in the vicinity. A downsampling operation is then applied to minimize the geometric and color information at the boundaries of various objects in indoor point cloud scenes. This approach simplifies the training and testing processes for the point cloud feature extraction module. Our experimental results demonstrate the efficacy of this method, with the network's performance significantly surpassing that of other related weakly supervised point cloud semantic segmentation approaches.

Furthermore, we conducted a comprehensive evaluation and ablation study of the proposed neural network through a series of experiments using the S3DIS dataset [10] and the ShapeNet dataset [25]. Comparative analysis with other relevant networks reveals that our network offers considerable improvements in terms of processing time and segmentation accuracy. This demonstrates its suitability for the task of semantic segmentation of point clouds in indoor environments.

Author Contributions: Conceptualization, Y. Miao; Methodology, Y. Miao, G. Ren and X. Zhang; Software, G. Ren and H. Liu; Validation, Y. Miao, X. Zhang and F. Liu; Formal analysis, Y. Miao and H. Liu; Investigation, Y. Miao, X. Zhang and F. Liu; Data curation, G. Ren, H. Liu and F. Liu; Writing original draft preparation, Y. Miao and X. Zhang; Writing—review and editing, Y. Miao and F. Liu; Visualization, G. Ren and H. Liu; Supervision, Y. Miao; Project administration, Y. Miao and F. Liu; Funding acquisition, Y. Miao. All authors have read and agreed to the published version of the manuscript.

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

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: S3DIS dataset—[http://buildingparser.stanford.edu].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Doula, A.; Gudelhofer, T.; Matviienko, A.; Muhlhauser, M.; Guinea, A.S. Pointcloudlab: An environment for 3d point cloud annotation with adapted visual aids and levels of immersion. In Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2023; pp. 11640-11646.
2. Zhang, R.; Chen, S.; Wang, X.; Zhang, Y. IPCONV: Convolution with multiple different kernels for point cloud semantic segmentation. *Remote Sensing*, 2023, 15(21): 5136.
3. Ballouch, Z.; Hajji, R.; Poux, F.; Kharroubi, A.; Billen, R. A prior level fusion approach for the semantic segmentation of 3D point clouds using deep learning. *Remote Sensing*, 2022, 14(14): 3415.
4. Liu, L.; Zhuang, Z.; Huang, S.; Xiao, X.; Xiang, T.; Chen, C.; Wang, J.; Tan, M. Cpcm: Contextual point cloud modeling for weakly-supervised point cloud semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023; pp. 18413-18422.
5. Ouassit, Y.; Ardchir, S.; Ghoumari, M.Y.E.; Azouazi, M. A brief survey on weakly supervised semantic segmentation. *International Journal of Online & Biomedical Engineering*, 2022, 18(10): 83-113.
6. Mao, J.; Shi, S.; Wang, X.; Li, H. 3D object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 2023, 131: 1909-1963.

7. Qian, R.; Lai, X.; Li, X. 3D object detection for autonomous driving: A survey. *Pattern Recognition*, 2022, 130: 108796.
8. Miao, Y.; Xiao, C. Geometric processing and shape modeling of 3d point-sampled models. *Beijing: Science Press*, 2014; pp.50-63.
9. Zhu, Y.; Mottaghi, R.; Kolve, E.; Lim, J.J.; Gupta, A.; Li, F.F.; Farhadi, A. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2017; pp. 3357-3364.
10. Armeni, I.; Sener, O.; Zamir, A. R.; Jiang, H.; Brilakis, I.; Fischer, M.; Savarese, S. 3D semantic parsing of large-scale indoor spaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016; pp. 1534–1543.
11. Kim, S.; Lee, S.; Hwang, D.; Lee, J.; Hwang, S. J.; Kim, H. J. Point cloud augmentation with weighted local transformations. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021; pp. 548-557.
12. Li, R.; Li, X.; Heng, P.A.; Fu, C.W. Pointaugment: an auto-augmentation framework for point cloud classification. In Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020; pp. 6378-6387.
13. Leng, Z.; Cheng, S.; Caine, B.; Wang, W.; Zhang, X.; Shlens, J.; Tan, M.; Anguelov, D. Pseudoaugment: Learning to use unlabeled data for data augmentation in point clouds. In Proceedings of European Conference on Computer Vision (ECCV), 2022; pp. 555-572.
14. Wang, R.; Yang, Y.; Tao, D. Art-point: Improving rotation robustness of point cloud classifiers via adversarial rotation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022; pp. 14371-14380.
15. Miao, Y.; Ren, G.; Wang, J.; Liu, F. Weakly supervised semantic segmentation for point cloud based on view-based adversarial training and self-attention fusion. *Computers & Graphics*, 2023, 116: 46-54.
16. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multiview convolutional neural networks for 3d shape recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015; pp. 945-953.
17. Zhou, Y.; Tuzel, O. Voxelnet: end-to-end learning for point cloud based 3d object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018; pp. 4490-4499.
18. Ni, P.; Li, X.; Xu, W.; Zhou, X.; Jiang, T.; Hu, W. Robust 3d semantic segmentation method based on multi-modal collaborative learning. *Remote Sensing*, 2024, 16(3): 453.
19. Nan, L.; Xie, K.; Sharf, A. A search-classify approach for cluttered indoor scene understanding. *ACM Transactions on Graphics*, 2012, 31(6): Article No. 137.
20. Li, Y.; Dai, A.; Guibas, L.; Nießner, M. Database assisted object retrieval for real-time 3d reconstruction. *Computer Graphics Forum*, 2015, 34: 435-446.
21. Shi, Y.; Long, P.; Xu, K.; Huang, H.; Xiong, Y. Data driven contextual modeling for 3d scene understanding. *Computers & Graphics*, 2016, 55: 55-67.
22. Mattausch, O.; Panozzo, D.; Mura, C.; Sorkine-Hornung, O.; Pajarola, R. Object detection and classification from large-scale cluttered indoor scans. *Computer Graphics Forum*, 2014, 33: 11-21.
23. Wu, Y.; Yan, Z.; Cai, S.; Li, G.; Han, X.; Cui, S. Pointmatch: a consistency training framework for weakly supervised semantic segmentation of 3d point clouds. *Computers & Graphics*, 2022, 116: 427436.
24. 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 Proceedings of the 17th European Conference on Computer Vision (ECCV), 2022; pp. 600-619.
25. Chang, A. X.; Funkhouser, T.; Guibas, L.; 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. arXiv, 2015, preprint arXiv: 1512.03012.
26. Mo, K.; Zhu, S.; Chang, A.X.; Yi, L.; Tripathi, S.; Guibas, L.J.; Su, H. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR), 2019; pp. 909-918.
27. Tchapmi, L.; Choy, C.; Armeni, I.; Gwak, J.; Savarese, S. Segcloud: Semantic segmentation of 3d point clouds. In Proceedings of IEEE International Conference on 3D Vision (3DV), 2017; pp. 537-547.

28. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of Advances in neural information processing systems (NeuIPS), 2017; pp. 5105-5114.
29. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. Pointcnn: Convolution on x-transformed points. In Proceedings of Advances in Neural Information Processing Systems (NeuIPS), 2018; pp. 828-838.
30. Landrieu, L.; Simonovsky, M. Large-scale point cloud semantic segmentation with superpoint graphs. In Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018; pp. 4558-4567.
31. Guo, M. H.; Cai, J. X.; Liu, Z. N.; Mu, T. J.; Martin, R. R.; Hu, S. M. PCT: Point cloud transformer. *Computational Visual Media*, 2021, 7: 187-199.
32. Jiang, L.; Zhao, H.; Liu, S.; Shen, X.; Fu, C.W.; Jia, J. Hierarchical point-edge interaction network for point cloud semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019; pp. 10433-10441.
33. Choy, C.; Gwak, J. Y.; Savarese, S. 4D spatio-temporal convnets: Minkowski convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019; pp. 3075-3084.
34. Thomas, H.; Qi, C.R.; Deschaud, J.E.; Marcotegui, B.; Goulette, F.; Guibas, L.J. Kpconv: Flexible and deformable convolution for point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019; pp. 6411-6420.
35. Laine, S.; Aila, T. Temporal ensembling for semi-supervised learning. arXiv 2016, preprint arXiv:1610.02242 .
36. Lu, T.; Wang, L.; Wu, G. Cga-net: Category guided aggregation for point cloud semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021; pp. 11693-11702.
37. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. Randla-net: Efficient semantic segmentation of large-scale point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020; pp. 11108-11117.
38. Liu, Z.; Hu, H.; Cao, Y.; Zhang, Z.; Tong, X. A closer look at local aggregation operators in point cloud analysis. In Proceedings of the 16th European Conference Computer Vision (ECCV), 2020; pp. 326-342.
39. Su, H.; Jampani, V.; Sun, D.; Maji, S.; Kalogerakis, E.; Yang, M.H.; Kautz, J. Splatnet: Sparse lattice networks for point cloud processing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018; pp. 2530-2539.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.