buildings

MDPI

Article

MSFA-Net: A Multiscale Feature Aggregation Network for Semantic Segmentation of Historical Building Point Clouds

Ruiju Zhang [1,2,3], Yaqian Xue [1], Jian Wang [1,4], Daixue Song [5,*], Jianghong Zhao [1,2,3] and Lei Pang [1]

[1] School of Geomantic and Urban Information, Beijing University of Civil Engineering and Architecture, Beijing 102616, China; zhangruiju@bucea.edu.cn (R.Z.); 2108570021125@stu.bucea.edu.cn (Y.X.); wangjian@bucea.edu.cn (J.W.); zhaojiangh@bucea.edu.cn (J.Z.); panglei@bucea.edu.cn (L.P.)
[2] Engineering Research Center of Representative Building and Architectural Heritage Database, Ministry of Education, Beijing 102616, China
[3] Beijing Key Laboratory for Architectural Heritage Fine Reconstruction & Health Monitoring, Beijing 102616, China
[4] Institute of Science and Technology Development, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
[5] Beijing Urban Construction Design & Development Group Co., Ltd., Beijing 100037, China
* Correspondence: tommy_song_beijing@163.com

Abstract: In recent years, research on the preservation of historical architecture has gained significant attention, where the effectiveness of semantic segmentation is particularly crucial for subsequent repair, protection, and 3D reconstruction. Given the sparse and uneven nature of large-scale historical building point cloud scenes, most semantic segmentation methods opt to sample representative subsets of points, often leading to the loss of key features and insufficient segmentation accuracy of architectural components. Moreover, the geometric feature information at the junctions of components is cluttered and dense, resulting in poor edge segmentation. Based on this, this paper proposes a unique semantic segmentation network design called MSFA-Net. To obtain multiscale features and suppress irrelevant information, a double attention aggregation module is first introduced. Then, to enhance the model's robustness and generalization capabilities, a contextual feature enhancement and edge interactive classifier module are proposed to train edge features and fuse the context data. Finally, to evaluate the performance of the proposed model, experiments were conducted on a self-curated ancient building dataset and the S3DIS dataset, achieving OA values of 95.2% and 88.7%, as well as mIoU values of 86.2% and 71.6%, respectively, further confirming the effectiveness and superiority of the proposed method.

Keywords: deep learning; historical building point cloud; MSFA-Net; semantic segmentation

Citation: Zhang, R.; Xue, Y.; Wang, J.; Song, D.; Zhao, J.; Pang, L. MSFA-Net: A Multiscale Feature Aggregation Network for Semantic Segmentation of Historical Building Point Clouds. *Buildings* **2024**, *14*, 1285. https://doi.org/10.3390/buildings14051285

Academic Editor: Yi-Kai Juan

Received: 24 March 2024
Revised: 16 April 2024
Accepted: 29 April 2024
Published: 1 May 2024

1. Introduction

The historical architecture of the Chinese nation is considered to be the oldest architectural system with the longest history, the longest existence, and the highest innovation, and has extremely high cultural heritage value. However, with the erosion of time, historic buildings are threatened with destruction and extinction. In recent years, the protection of cultural heritage has been continuously strengthened at the national level and a number of policy documents have been implemented successively. This makes the protection and repair of historical buildings in our country increasingly important and urgent.

With the continuous deepening of 3D digital technology in large-scale cultural heritage protection research [1], point cloud semantic segmentation has become an important direction for remote sensing applications. Different from traditional pixels in two-dimensional images, point clouds have more detailed depth information and provide a large amount of valuable information to describe the real world [2]. However, point cloud data also have shortcomings. Although point cloud data contains three-dimensional coordinates and various additional attributes with high precision, high resolution, and high dimension, it

cannot directly offer information at the semantic level. These problems make it difficult for cultural heritage experts to directly use point cloud data. Therefore, as a basic link, point cloud segmentation of historical buildings has important research significance.

Nowadays, point cloud semantic segmentation has become the basic technology for three-dimensional scene understanding, and researchers have conducted in-depth exploration on it, but its unique disordered and unstructured characteristics make it difficult to obtain precise and effective semantic segmentation outcomes, which is extremely challenging [3–5]. In the early stage, people were committed to using the traditional point cloud semantic segmentation method, which divides the point cloud data into different surface regions according to the feature attributes of the point cloud. Its algorithms are divided into four methods: edge based, region-growing based [6], model-fitting based [7] and clustering based. Each algorithm has its own unique advantages and disadvantages in point cloud semantic segmentation, as well as characteristics which are applicable to different scenarios [8]. Although the traditional point cloud segmentation method performs well in man-made structures with regular geometric shapes and runs faster, there are still some limitations in large-scale historical buildings. For example, most historical buildings consist of a large number of components with irregular shapes, so it is difficult to select suitable geometric models to fit objects. Only relatively rough segmentation results can be obtained.

With the deepening of research, deep learning technology has propelled amazing advancements in point cloud semantic segmentation. Numerous methods for semantic segmentation of point clouds using deep learning techniques have emerged in recent years [9]. In contrast to the conventional point cloud semantic segmentation technique, the deep learning-based model technology not only has higher performance of multiscale spatial three-dimensional information but also has different granularity levels of semantic information, include partial segmentation, semantic segmentation, and instance segmentation. They can be categorized into three types based on various point cloud extraction techniques: voxel-based methods [10,11], projector-based methods [12,13], and point-based methods [14]. The projection-based method and voxel-based method have high computational costs, and it is easy to cause semantic feature or spatial position loss in the process of projection or voxelization. In order to address this issue, researchers constructed a network for collecting features from point clouds without requiring data transformation processes [15–17], which can directly consume irregular three-dimensional point cloud data, reduce the limitation of point cloud characteristics, and make full use of point cloud geometry information to improve the interpretation ability of three-dimensional point cloud scenes. PointNet [18], as a pioneer of deep learning, first provides a network architecture that directly handles the original point cloud. Therefore, many scholars have proposed improved networks based on it, but most of the methods are limited to the input of minuscule three-dimensional point clouds into the network and cannot be directly extended to larger scenarios. Subsequently, Hu et al. [19] proposed a RandLA-Net network model with better performance in large-scale scenarios. It chooses random sampling instead of the widely used remote point sampling method and extracts geometric features through the local feature aggregation module, which reduces the network complexity and effectively retains the geometric details. On this basis, many methods of local feature aggregation have recently been presented. As an illustration, SCF [20] introduces spatial representations that are not affected by Z-axis rotation, LACV-Net [21] uses the neighborhood feature as the offset and converges to the centroid feature, which reduces the local perceptual ambiguity through its similarity, and DGFA-Net [22] has an expansion graph characteristic aggregate structure.

According to the above analysis, although point-based methods have obtained good accuracy in semantic segmentation, they are rarely used for historical architecture in China. When dealing with large-scale ancient architectural scenes, these methods cannot sufficiently capture both local and global information, especially when faced with uniquely structured historical architectural components. Overemphasis on local features may ne-

glect the spatial geometric structure information of the point cloud. Therefore, the main contributions of this study are as follows:

(1) This paper proposes a unique semantic segmentation network named MSFA-Net. It designs a double attention aggregation (DAA) module, which consists of a bidirectional adaptive pooling (BAP) block and a multiscale attention aggregation (MSAA) block. Through the combination of two different attention mechanisms, it can obtain multiscale information features of the target in the sampling process and reduce redundant information.

(2) This paper proposes a contextual feature enhancement (CFE) module, which enhances the connection between the model context by fusing the local global features across the encoding and decoding layers and fully considers the semantic gap between neighboring features.

(3) This paper proposes an edge interactive classifier (EIC), which introduces the features of each point into the edge interactive classifier to obtain the edge features of each point. Through the information transfer between nodes, it better performs label prediction, making it possible to smoothly segment the edges of objects.

2. Materials and Methods

2.1. MSFA-Net Model Construction

This section outlines the proposed network's detailed design, which follows the encoder–decoder structure and is depicted in Figure 1 of this text. The input point cloud is introduced into the coding layer, which includes a downsampling operation and a dual attention aggregation module (DAA). In view of the efficiency of random sampling in dealing with large-scale point cloud data, this study adopts this method to simplify point cloud data. Both upsampling and MLP operations are present in every decoding layer. Then, we add a CFE module between the encoders and decoders, allowing features to be transferred interactively across the encoding and decoding layers, fusing the local features the encoding and decoding layers had previously gathered to improve context linkage. Finally, an EIC module was designed to obtain the edge features of each point and perform better label prediction through information transmission between nodes. The label with the highest score is used to determine the semantic segmentation outcome.

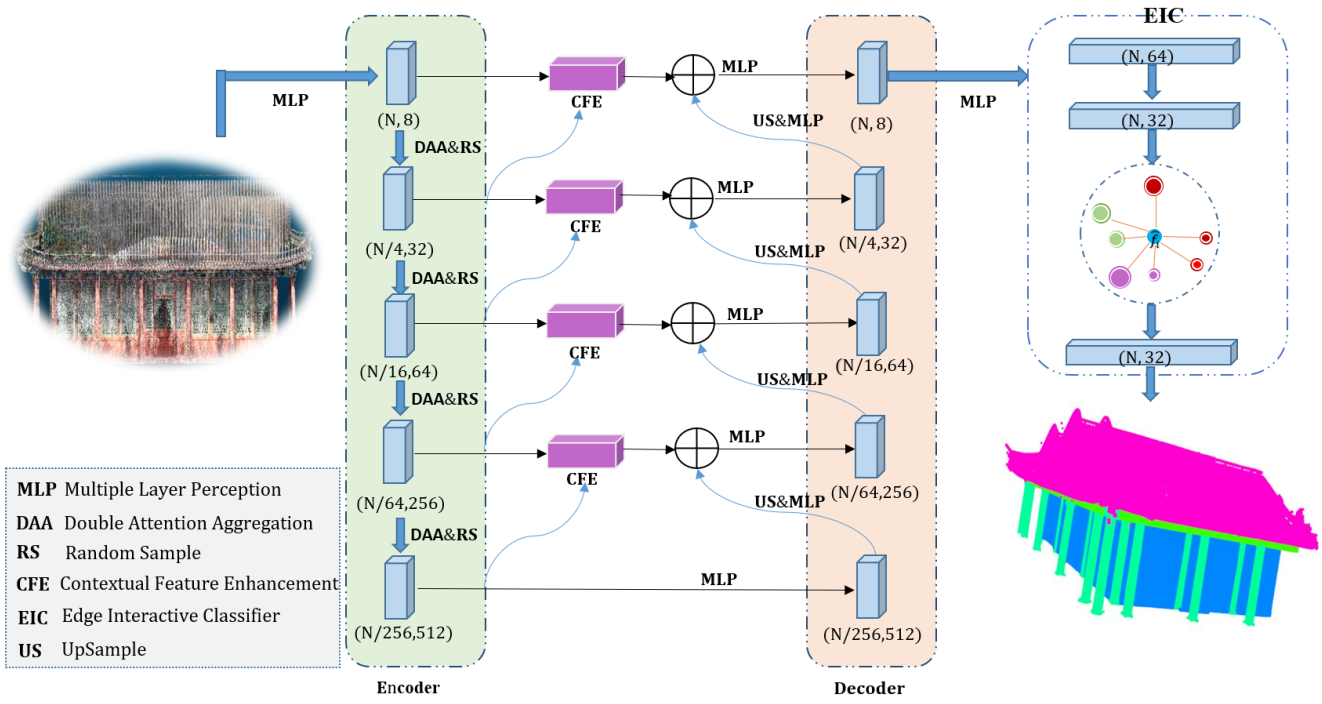

Figure 1. MSFA-Net network structure.

2.2. The Structure of the Double Attention Aggregation (DAA) Module

Figure 2 depicts the DAA module's detailed structure, and the input point information comprises the previously learned spatial and feature information. Point coding blocks are built using spatial information, and weighted learning is accomplished via bidirectional adaptive pooling (BAP) and multiscale attention aggregation (MSAA) modules. The next coding layer receives all the acquired features and processes them.

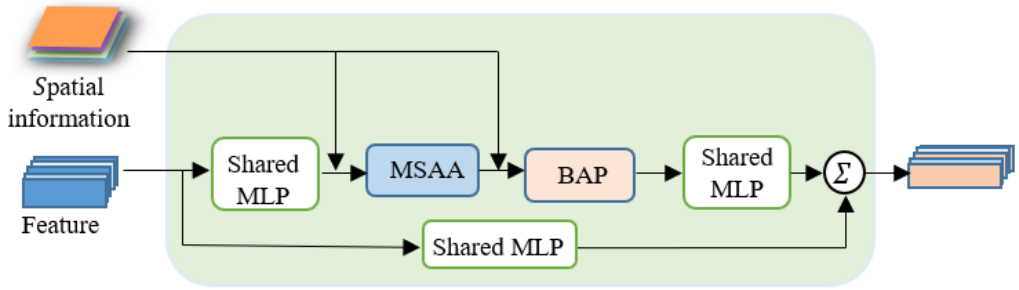

Figure 2. DAA module structure diagram.

The numerical symbols utilized in Figure 2, as well as the other number symbols depicted in the module schematic diagrams shown in Figures 2–6, are detailed in Table 1.

Table 1. List of abbreviations.

Symbol	Explanation
ⓚ	K nearest neighbor
Ⓢ	Sum
⊖	Subtract
⊕	Concatenation
Ⓢ	Softmax
⊙	Hadamard product
BN	Batch normalization with ReLU activation
Ⓣ	Transpose

2.2.1. Multiscale Attention Aggregation Module

The present study introduces a multiscale attention aggregation module, which extracts feature information from different scales. This module computes corresponding attention scores based on neighbor information and utilizes them to weight the neighbor features, resulting in the acquisition of the final aggregate feature vector. Figure 3 depicts our multiscale attention aggregation module.

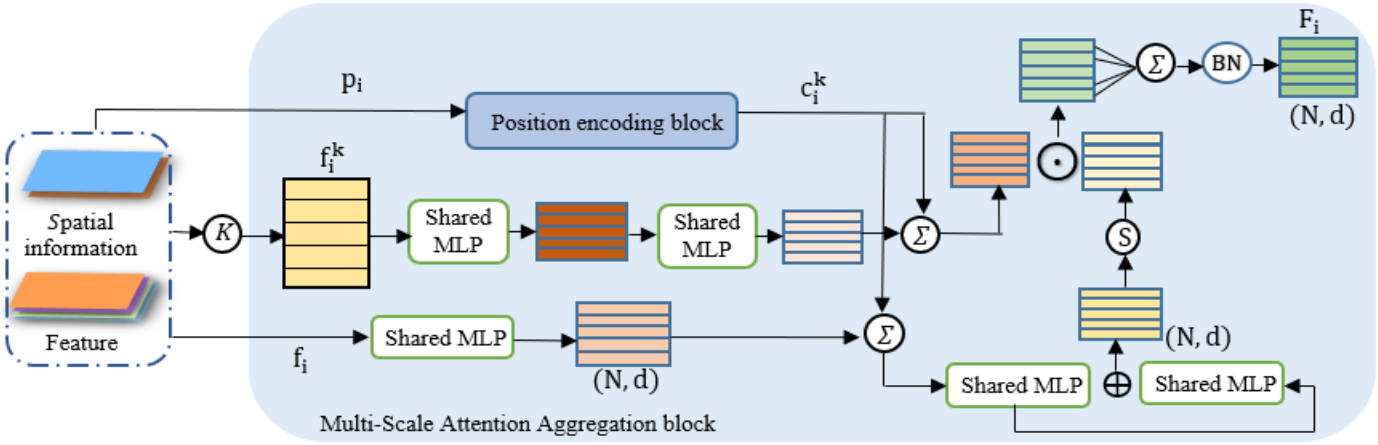

Figure 3. Multiscale attention aggregation module.

The central point and nearby points are first subjected to feature extraction in this module; then, position encoding [19] is added to both the mapping function ζ and the conversion function γ to enhance the ability to express features. Let f_i be the characteristic of the central point p_i, and f_i^k be the characteristic of the points that surround it. The following is the formula for our multiscale attention aggregation module:

$$F_i = \sum_{k=1}^{K} softmax\left(\zeta\left(\alpha(f_i) + c_i^k\right)\right) \odot \left(\gamma\left(f_i^k\right) + c_i^k\right) \quad (1)$$

In this module, by utilizing feature information at different scales, the receptive field is expanded to increase the perceptual ability of the point cloud. Not only can it better capture the geometric features of the characteristics, but it can also dynamically adjust the importance of the features according to their significance, focus on minute details, and reduce the problem of mis-segmentation.

2.2.2. Bidirectional Adaptive Pooling Module

Attention mechanisms have been widely used in computer vision technology and have been added to various segmentation tasks. A bidirectional adaptive pooling block is created, combining spatial information in response to the drawback that it is easy to unintentionally lose crucial information in random sampling tasks. This block can not only better capture forward and backward information but also enhance the model's capacity to recognize important features. Figure 4 depicts our bidirectional adaptive pooling structure.

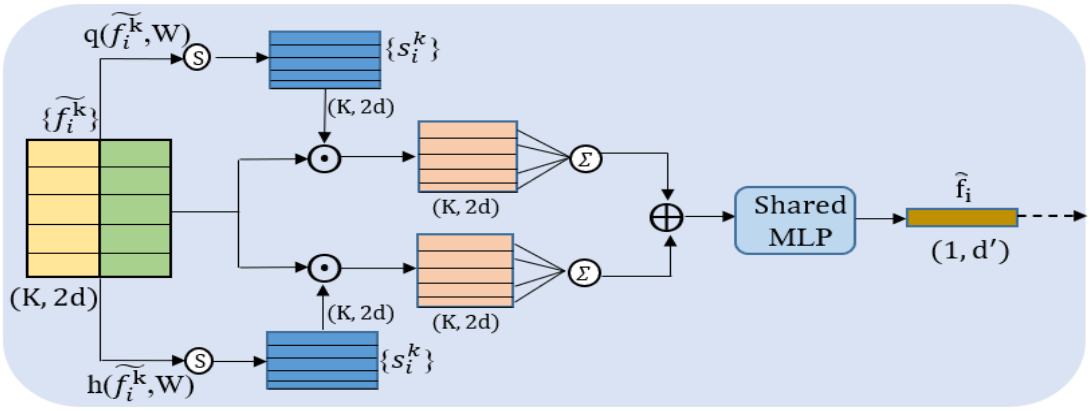

Figure 4. Bidirectional adaptive pooling module.

To obtain the aggregate feature set $\hat{F}_i = \left\{\widetilde{f}_i^k\right\}_{k=1}^{K}$, a shared function q (.) is designed to learn the aggregation of local features centered on the current point. Every feature is given a distinct attention score in its neighbor points by the shared MLP's learnable weight W_i. It is used to gauge how similar the present points are to those further away and to gauge the significance of screening features. Then, a shared function h(.) is designed to learn the weights of points that are closer to the current point. Each function includes a softmax and a shared MLP. The following is a definition of this form:

$$S_i^k = q\left(\widetilde{f}_i^k, W_i\right), \tag{2}$$

$$S_i^k = g\left(\widetilde{f}_i^k, W_i\right) \tag{3}$$

where W_i shares the MLP's learnable weight.

Then, the weighted summation is used to determine the attention scores for both forward and backward learning, which can be thought of as soft masks that automatically select important features. The weighted feature summation is performed, and the formula is as follows:

$$\hat{f}_i = \sum_{k=1}^{K}\left(\widetilde{f}_i^k \cdot S_i^k\right) \tag{4}$$

The final information feature vector $\hat{f}_i$ is created by combining the forward and backward weighted sums after each weighted sum is collected. Our method represents the features more accurately and comprehensively because it takes into account how each point is related to its neighbors.

2.3. The Structure of the Contextual Feature Enhancement (CFE) Module

To enhance the relationship between model contexts, this study first creates a module based on contextual feature enhancement that combines local and global characteristics from the encoder and decoder. Figure 5 depicts its structure. From the figure, it can be seen that the feature representation corresponding to the decoder layer of F_e^l is denoted as F_d^l. For F_d^l in the decoder, it is enhanced using F_e^l and F_e^{l+1} from the encoder.

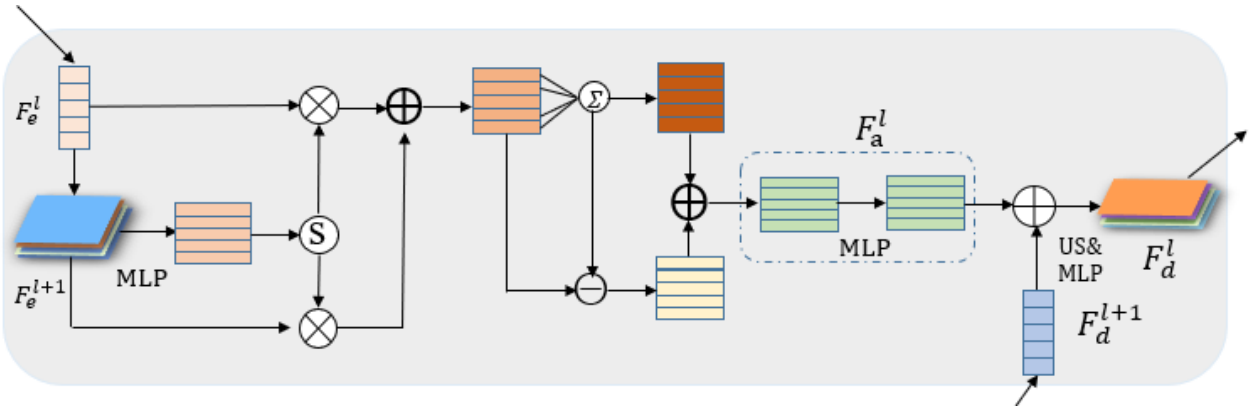

Figure 5. Contextual feature enhancement module.

First, to increase computational efficiency, we compress the feature map of the encoder in the preceding layer. Then, the nearest neighbor interpolation is used to interpolate the feature graph, and the convolution operation is used to map the feature, and the attention weight is obtained by normalizing the sigmoid function. Finally, the attention weight is employed to weigh the last encoder feature map and the current feature map to accomplish the feature map fusion. The following is the formula:

$$F_a^{\mathrm{l}} = mF_e^{\mathrm{l}} + M\left(F_e^{\mathrm{l}} - mF_e^{\mathrm{l}}\right) \tag{5}$$

where M stands for MLP and m stands for average.

Based on the acquired feature map, feature enhancement is performed on both the global and local features. Contextual associations within the input data are identified, enabling the model to prioritize crucial feature information. The following is the formula:

$$F_c^{\mathrm{l}} = F_e^{\mathrm{l}} \odot F_a^{\mathrm{l}} + F_e^{\mathrm{l}+1} \odot \left(1 - \mathrm{US}\left(\mathrm{M}\left(F_e^{\mathrm{l}+1}\right)\right)\right) \tag{6}$$

In this process, the attention map weights the feature maps at different positions, allowing the model to focus more on the regions related to the current task. By weighting the feature diagram of the preceding layer encoder, the model may be made to target high-level characteristics more, to produce more precise models of important regions, enhancing the performance of the model. Additionally, this weighting method can also play a regularization role to avoid the occurrence of overfitting problems.

2.4. The Structure of the Edge Interactive Classifier (EIC) Module

In previous studies, most classifiers use the full-connection layer and dropout layer to predict their classes point by point, but the nonactive functions that make up the full-connection layer easily cause inconsistent neighbors during prediction. Therefore, an edge interaction module is proposed to obtain edge features by exchanging information between adjacent nodes. Figure 6 depicts the edge interactive classifier module [23].

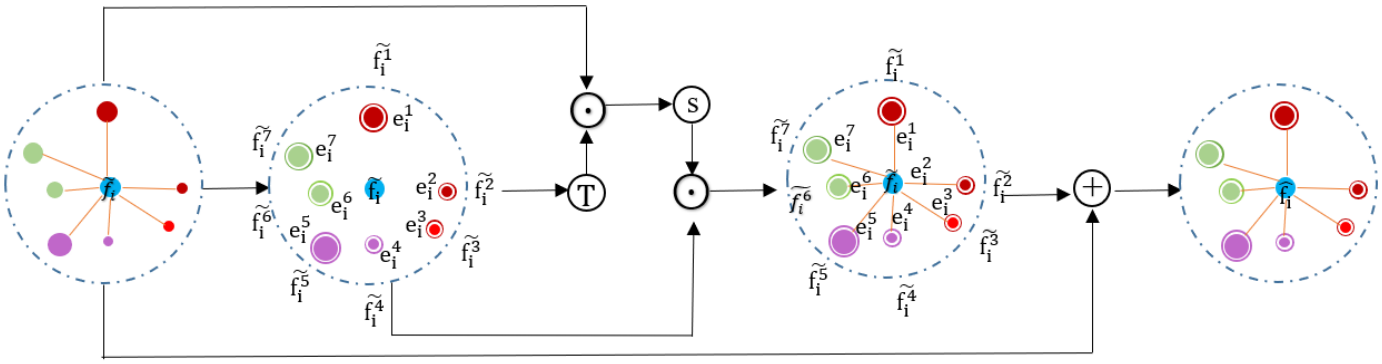

Figure 6. Edge interactive classifier module.

The point cloud is represented by $\widetilde{\mathrm{P}} = \left\{\mathrm{x_n}, \mathrm{y_n}, \mathrm{z_n}, \widetilde{\mathrm{f_n}}\right\}_{\mathrm{n=1}}^{\mathrm{N}}$, where x_n, y_n, z_n stand for the NTH point's coordinates, and $\widetilde{f_n} \in R^D$ denotes the feature output of the local aggregation module's preceding layer, by which the feature graph of the neighbor node is aggregated. For vertex $\widetilde{f_i}$, the nearest k vertex feature $\left\{\widetilde{f_i^k}\right\}_{k=1}^{K}$ is obtained and obtain edge features by echoing with the feature map, where the edge feature is $\mathrm{e_i^j} = \mathrm{ReLU}\left(\mathrm{W} \cdot \widetilde{f_i^j}\right)$. Where W is the learnable weight and ReLU is the ReLU activation function. The edge feature between the ith point and its jth neighbor is known as $\mathrm{e_i^j}$.

Then, to determine the relationship between each node and its neighbors, compute the attention matrix for the original feature and the edge feature:

$$a_j^i = \left(e_i^j\right)^T \cdot e_k^j \tag{7}$$

$$f_i' = \sum_j f_j^i \cdot E_j \tag{8}$$

where f_i' represents the aggregation feature of node i, E_j is the neighbor node j's feature matrix, and the attention coefficient f_j^i represents the relationship between node i and its neighbor node j. Finally, perform maximum pooling on it, select the most significant features, and combine them with the output of the preceding convolution layer to create fusion features. In order to improve contextual information and capture edge-to-edge interaction features, this module is included in the final two MLPs.

3. Experiment

3.1. Experimental Platform

This experiment was run on NVIDIA GeForce RTX 3080/PCIe/SSE2, Intel® Core^TMi9-10900K CPU @ 3.70 GHz *20 in Ubuntu 18.04 environment. We adopted TensorFlow, which is implemented in Python 3.6, as the framework for our algorithm. During the training phase, the network selects eight batch sizes and performs 100 iterations epochs of the calculation. Concurrently, the starting learning rate is fixed at 0.01, and after each epoch, it is reduced by 10%. For gradient descent optimization, the network optimizer makes use of Adam [24].

3.2. Dataset

Two historic buildings were selected as study areas for this study. The first research area is located in Beijing Olympic Park, Chaoyang City, which dates back to the Xuande period of the Ming Dynasty (1426–1435) and has been continuously maintained until now. It is a religious building with a long history and cultural heritage. The other study area, which is different from the religious buildings, is a Chinese classical building, built in the Qing Dynasty (1908), located in the Xicheng district of Beijing, with extremely important historical significance.

In order to obtain complete point cloud data of historical buildings in real scenes, we use 3D laser scanning technology to obtain historical building data containing complete internal and external structures, then preprocess them using a range of data processing software, such as registration, filtering, sampling, and annotation. To objectively assess the semantic segmentation network model's capacity for generalization on historical building data, based on the symmetry of historical architectural scenes and their structures, we divided each building into left and right point clouds along the central axis, and marked a total of 6 semantic categories: roof, column, forehead, door, window, and wall. This segmentation method not only helps the neural network to better learn the structural characteristics of historical building components, but also adopts the four-fold cross-validation method in the subsequent steps, so as to scientifically evaluate the model's capacity for generalization. The manual annotation results of the point clouds are shown in the Figure 7:

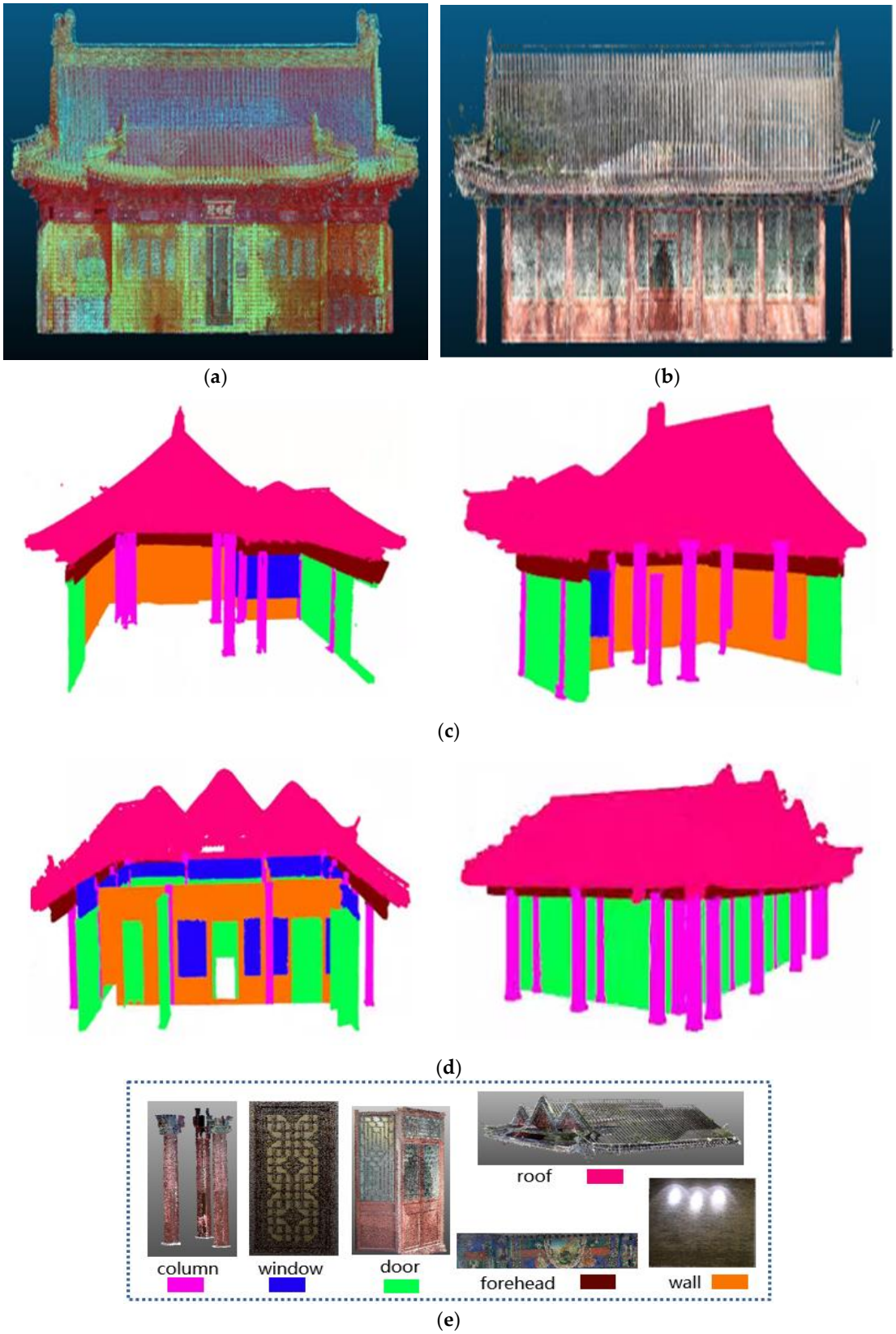

Figure 7. The experimental area manual annotation results. (**a**,**b**) The point cloud data of the experimental area scene, (**c**) the annotation result of the point cloud data of experimental area 1 (left and right), (**d**) the annotation result of the point cloud data of experimental area 2 (left and right), and (**e**) the wood component division of the ancient building dataset.

3.3. Evaluation Indicators

To be able to accurately reflect the performance of the presented approach in this study, the study chose the classical evaluation criteria to validate it, namely, overall accuracy (OA), mean class accuracy (mAcc), and mean intersection over union (mIoU). The calculation formula of each part is as follows:

$$\mathrm{OA} = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \tag{9}$$

$$\mathrm{m}Acc = \frac{i}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{10}$$

$$\mathrm{m}IoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{11}$$

Among them, k denotes the number of labels existing in the used dataset, p_{ij} represents the number of points where samples originally classified as class i were misclassified as class j, p_{ii} represents the number of points where samples originally classified as class i were correctly classified as class i, and p_{ji} represents the number of points where a sample originally classified as class j was misclassified as class i.

4. Results and Discussion

This section includes various experiments designed to test the performance of the proposed MSFA-Net model. The study selected a self-curated historical building dataset and a typical large-scale benchmark set, S3DIS [25], for evaluation. Under the same conditions, it can be intuitively found that the MSFA-Net model has high accuracy and is more suitable for semantic segmentation of point clouds of historical buildings by comparing the MSFA-Net model with other semantic segmentation models. In addition, this article also conducted ablation experiments to assess the modules' efficacy of the modules from an objective perspective. It is further proved that our suggested MSFA-Net model can address the issues of inaccurate division of historical building components and unclear boundary position to a large extent.

4.1. Comparison with Other Methods

In this paper, the proposed method is evaluated using four-fold cross-validation on a self-curated historical building dataset and is quantitatively compared with semantic segmentation methods that have achieved good results in recent years, including Rand-LA-Net, BAF-LA [26], DLA-Net [27], and SCF-Net. Table 2 displays the outcomes of the experiment. Obviously, our proposed network is superior to the above networks in OA (95.2%), mAcc (92.5%) and mIoU (86.2%) and has relative improvements of 0.7%, 1.4%, and 1.6% compared with RandLA-Net in three quantitative aspects, respectively.

Table 2. Segmentation results of different networks on the historical building datasets. The class metric is IoU (%).

	OA (%)	mAcc (%)	mIoU (%)	Column	Door	Forehead	Wall	Roof	Window
RandLA-Net	94.5	91.1	84.6	80.0	88.1	83.7	85.8	98.1	71.9
BAF-LAC	93.4	90.2	82.4	79.1	85.2	83.4	80.3	97.1	69.3
DLA-Net	93.8	91.7	84.2	85.8	86.0	82.8	80.0	96.6	76.9
SCF-Net	95.0	92.0	85.0	81.9	86.9	84.5	84.8	98.3	73.8
Our	95.2	92.5	86.2	84.9	88.9	84.7	84.8	98.3	75.8

As shown in Figure 8, the forecast results for the several methods covered in Table 2 are shown, where black circles, brown ellipses, and red boxes all show parts of the different networks that are poorly segmented or where boundary segmentation is not obvious.

The previous semantic segmentation method is effective in the segmentation of historical building scenes, but there are still some problems, such as incomplete segmentation of columnar components and unclear segmentation of architectural boundaries. From the outside of area 3 in the Figure 8a, it can be observed that compared to the real annotation, the RandLA-Net model produces incomplete cylindrical segmentation in the areas marked by brown ellipses, with significant omissions. The BAF-LAC model shows fewer missing parts. Although the DLA Net and SCF Net models did not exhibit undersegmentation, both exhibited oversegmentation. In the area marked by black circles, all models exhibit varying degrees of oversegmentation. Both RandLA Net and DLA Net models have misclassified the edges of the objects marked in red boxes, while BAF-LAC and SCF Net models have no large area errors, but the segmentation boundaries are not smooth enough. In summary, RandLA Net performs poorly in terms of columnar integrity and edge segmentation results, with BAF-LAC, DLA Net, and SCF Net showing an increasing effect. However, the MSFA-Net proposed in this paper performs better overall than other networks in the selected regions. Similarly, the last row of Figure 8c depicts area 1, which, like the exterior of area 3, features both three-dimensional and planar components. The brown oval mark represents incomplete segmentation, the black circle mark represents undersegmentation and oversegmentation, and the red box mark represents unclear edge segmentation. As can be seen in the figure, the MSFA-Net model has superior segmentation performance at the integrity and edges of column and planar components, while other networks have obvious missegmentation problems at the boundaries of different components. However, Figure 8b shows the interior of area 3, which mostly consists of flat categories such as doors, windows, beams, and walls. The black box indicates the location of misclassification. RandLA Net, BAF LAC, DLA Net, and SCF Net networks all have a large area of misclassification. It is clear that the MSFA-Net proposed in this article has a smaller area of misclassification compared to other networks, making it more suitable for point cloud semantic segmentation tasks in large-scale historical architectural scenes.

However, misclassification is inevitable, and the problem of unclear segmentation of similar scenes is common in all networks. It is still necessary to find improvement strategies to improve the semantic segmentation effect. In the research of point cloud semantic segmentation, the inherent noise, outliers, and uneven sampling density of raw data pose challenges to preprocessing algorithms. Currently, there is no preprocessing method that can completely eliminate these adverse effects. In addition, the limitations of the model structure, insufficient feature extraction, and insufficient generalization ability further affect the overall segmentation performance of the model. In summary, improving the above issues can help reduce the occurrence of unclear segmentation.

To objectively demonstrate the effectiveness of the proposed method, this study selected a typical large-scale indoor point cloud dataset S3DIS for validation. The S3DIS dataset is a semantically annotated indoor scene cloud dataset that contains 6 different indoor regions, producing 11 scenes with a total of 271 rooms. Each room is a medium-sized point cloud, which is divided into 13 categories and marked with its own label for each point. We compared the MSFA-Net model with seven classic methods, including PointNet, SPG [28], PointCNN [29], PointWeb [30], ShellNet [31], KPConv, and RandLA-Net, and conducted six cross-validations to evaluate the proposed method. The experiment uses mIoU, mAcc, and OA as standard metrics. Table 3 displays the outcomes of the experiment, which summarizes our network's quantitative results on the S3DIS dataset in comparison to other advanced underlying networks. The results show that our proposed network achieved excellent performance in OA, mAcc, and mIoU, which are 0.7%, 0.6%, and 1.6% better than RandLA-Net, respectively, and achieved excellent performance in the following three categories: beam, board, and clutter.

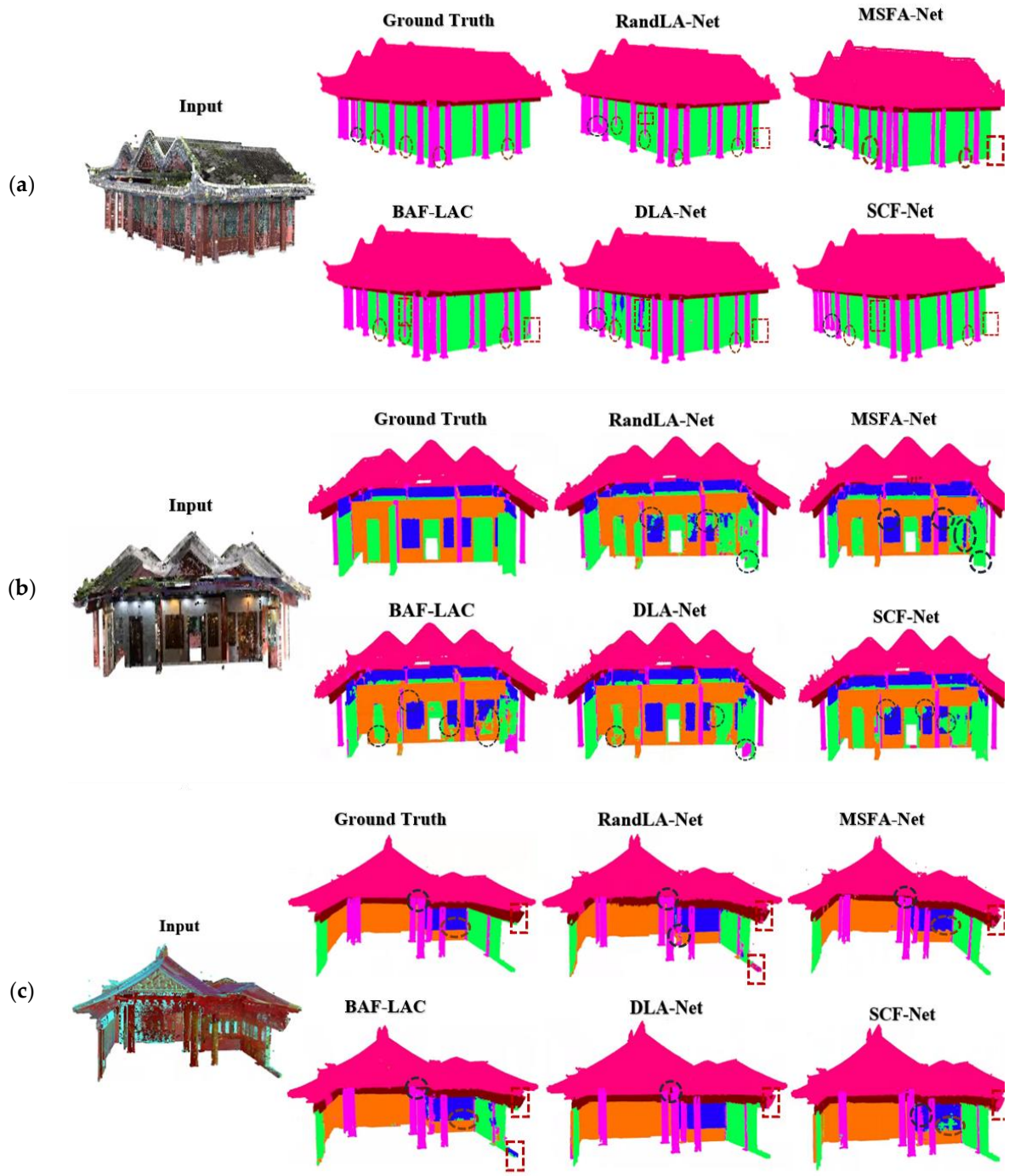

Figure 8. Comparison of segmentation effects on historical building datasets. In the figure, from top to bottom, the three areas are (**a**) outside area 3, (**b**) inside area 3, and (**c**) area 1.

To more clearly show the different effects of the network, this article intuitively compared the semantic segmentation results of RandLA-Net and MSFA-Net implementation on the S3DIS dataset, and the effect is shown in Figure 9. Figure 9 shows visualizations of three typical indoor scenarios: offices, conference rooms, and hallways. It is evident from the quantitative and qualitative results that our suggested MSFA-Net is capable of more correctly and fluidly segmenting object boundaries. However, semantic segmentation inevitably leads to misclassification, as shown in Figure 9, where a table in the conference room is misclassified.

Table 3. Segmentation results of different networks on S3DIS. The class metric is IoU (%).

	OA (%)	mAcc (%)	mIoU (%)	Ceiling	Floor	Wall	Beam	Column	Window	Door	Table	Chair	Sofa	Bookcase	Board	Cluster
PointNet	78.6	66.2	47.6	88.0	88.7	69.3	42.4	23.1	47.5	51.6	51.4	42.0	9.6	38.2	29.4	35.2
SPG	86.4	73.0	62.1	89.9	95.1	76.4	62.8	47.1	55.3	68.4	73.5	69.2	63.2	45.9	8.7	52.9
PointCNN	88.1	75.6	65.4	94.8	97.3	75.8	63.3	51.7	58.4	57.2	71.6	69.1	39.1	61.2	52.2	58.6
PointWeb	87.3	76.2	66.7	93.5	94.2	80.8	52.4	41.3	64.9	68.1	71.4	67.1	50.3	62.7	62.2	58.5
ShellNet	87.1	-	66.8	90.2	93.6	79.9	60.4	44.1	64.9	52.9	71.6	84.7	53.8	64.6	48.6	59.4
KPConv	-	79.1	70.6	93.6	92.4	83.1	63.9	54.3	66.1	76.6	57.8	64.0	69.3	74.9	61.3	60.3
RandLA-Net	88.0	82.0	70.0	93.1	96.1	80.6	62.4	48.0	64.4	69.4	69.4	76.4	60.0	64.2	65.9	60.1
Our	88.7	82.6	71.6	92.8	97.0	81.7	64.0	53.8	64.1	70.8	72.5	81.9	61.5	64.5	66.2	60.6

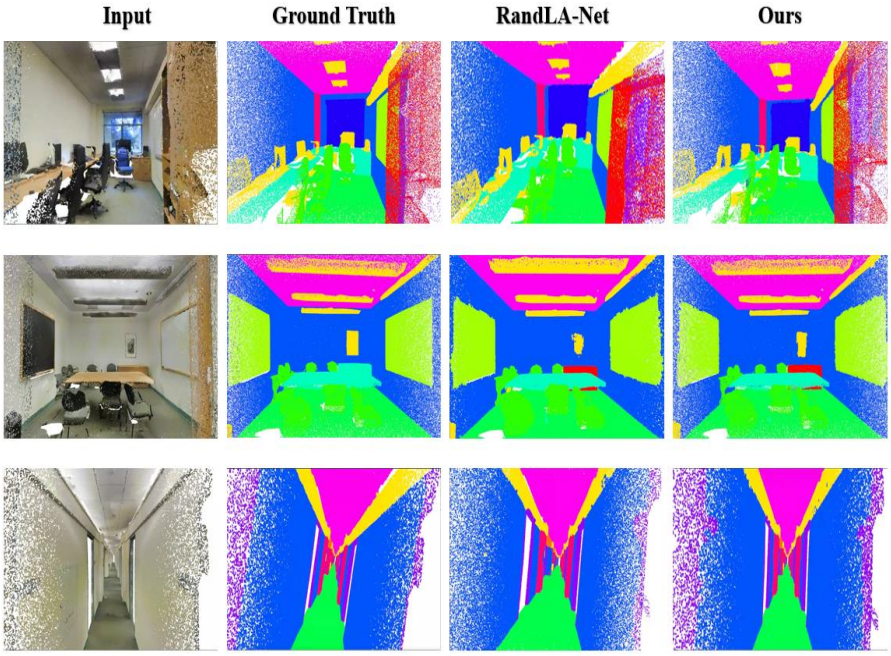

Figure 9. S3DIS data set segmentation effect.

4.2. Ablation Study

The validity of MSFA-Net is evidenced by the results of our experiments on the two datasets mentioned above. To better understand the model, the next step is to conduct ablation research, and all ablation networks were trained on homemade datasets of historical buildings. The study used standard 4-fold cross-validation to evaluate the ablation network and used OA, mIoU, and mAcc as metrics. We performed experiments on various combinations of these modules to evaluate the efficacy of the suggested modules quantitatively. As shown in Table 4, we first removed all modules and then added BAP, MSAA, CFE, and EIC into the model in turn. The table not only shows the individual effects of each module but also visually demonstrates that the performance gradually improves as more modules are added.

Table 4. Results of ablation studies on a dataset of historical buildings. The class metric is IoU (%).

Model Name	Modules				Evaluation Index		
	BAP	MSAA	CFE	EIC	OA (%)	mAcc (%)	mIoU (%)
None					87.8	81.3	71.1
BAP	√				94.1	89.5	83.0
MSAA		√			94.8	92.1	85.3
CFE			√		89.0	79.9	71.5
EIC				√	88.8	80.8	71.7
BAP + MSAA(DAA)	√	√			94.9	92.5	85.8
BAP + MSAA + CFE	√	√	√		95.0	92.4	86.0
Our	√	√	√	√	95.2	92.5	86.2

4.2.1. Ablation Experiment of DAA Module

(1) Ablation experiment of BAP module

This section confirms the efficacy of bidirectional adaptive pooling (BAP) blocks from several angles, including eliminating all and only positional coding blocks, replacing them with maximum pooling or average pooling, and using one-way learning feature information. The experiment was carried out on the historical building dataset, and the test results were carried out on area 2, as shown in Figure 10.

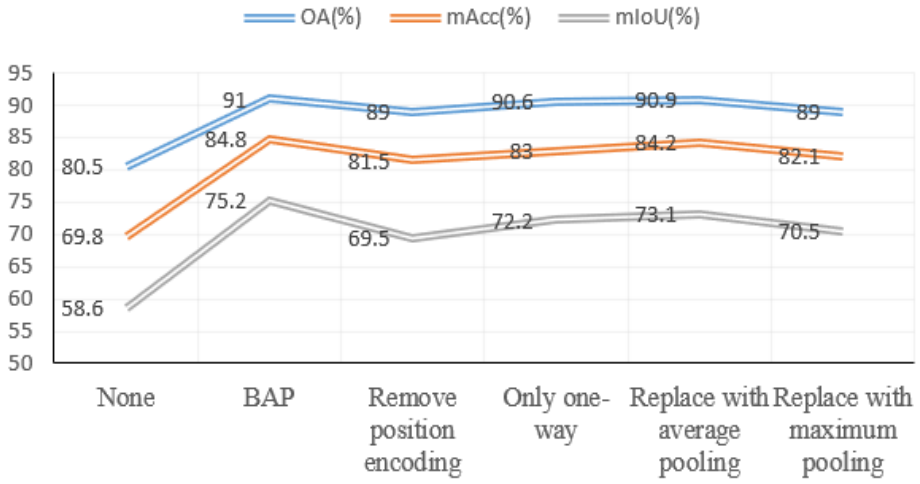

Figure 10. Results of ablation studies on the BAP module.

Experimental results have shown that the bidirectional adaptive pooling module achieves optimal performance in various accuracy indicators, indicating that the BAP module helps to enhance the model's ability to recognize key features and solve the problem of accidental loss of key information points during random sampling. The results also prove that the pooling method we used is better than average pooling and maximum pooling, and position encoding blocks play an important role.

(2) Ablation experiment of the MSAA module

To demonstrate the advantages of the MSAA over other attention mechanisms, experiments were conducted on an empty frame system. The MSAA is replaced with the dual-distance attentive pooling (DDAP) [20] module between the same network layer, and operations such as removing position encoding and replacing pooling are tested. Figure 11 displays the outcomes of applying various attention mechanisms to the model for the experimental analysis in the area 2 dataset.

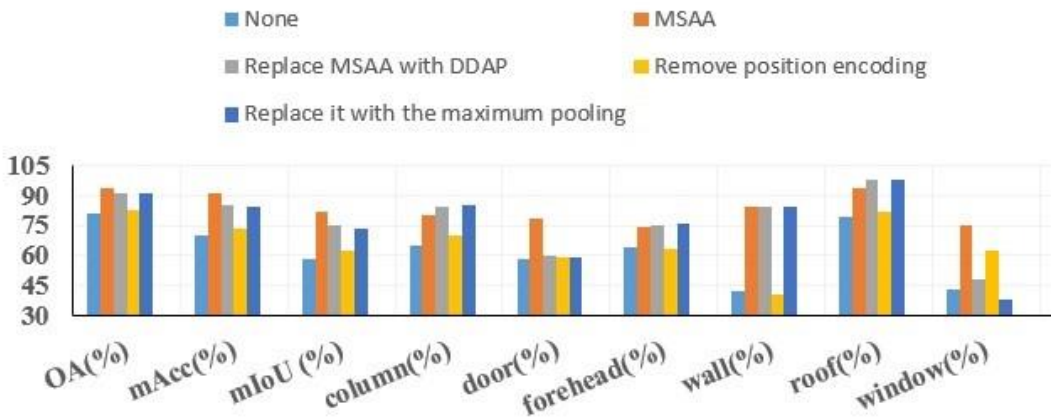

Figure 11. Ablation experiment of the MSAA module.

From the first three columns of the bar chart, it can be seen that the MSAA module proposed in this study has the highest accuracy in the three evaluation indicators, which clearly indicates that the MSAA proposed in this paper is more effective than the no attention mechanism and DDAP. As a novel feature enhancement module, DDAP adopts geometric distance and feature distance automatic learning methods for feature learning, which indirectly verifies that MSAA blocks can more fully learn feature information at different scales, suppress useless information, and fully mine local neighborhood information. Additionally, it is inferred from the figure that the performance of MSFA-Net will decline if there is no location-coding block in the MSAA block or if it is replaced by maximum pooling.

(3) DAA comparative experimental results and analysis

By conducting comparative experiments to verify the performance of using the DAA module to improve the RandLA Net model, this study trains the optimized RandLA Net model and the original RandLA Net model separately on the historical building dataset. The evaluation indicators and parameter settings used are as shown in the previous text.

Figure 12 clearly shows the visualization results of the comparative experiment of the RandLA Net model before and after improvement. The optimized model effectively overcomes the problem of insufficient feature extraction. It can be seen that each wooden component has a higher improvement in the integrity of the results compared to the original network segmentation, but it can also be seen that the final semantic segmentation effect is poor at the edges of the components, and further optimization is needed.

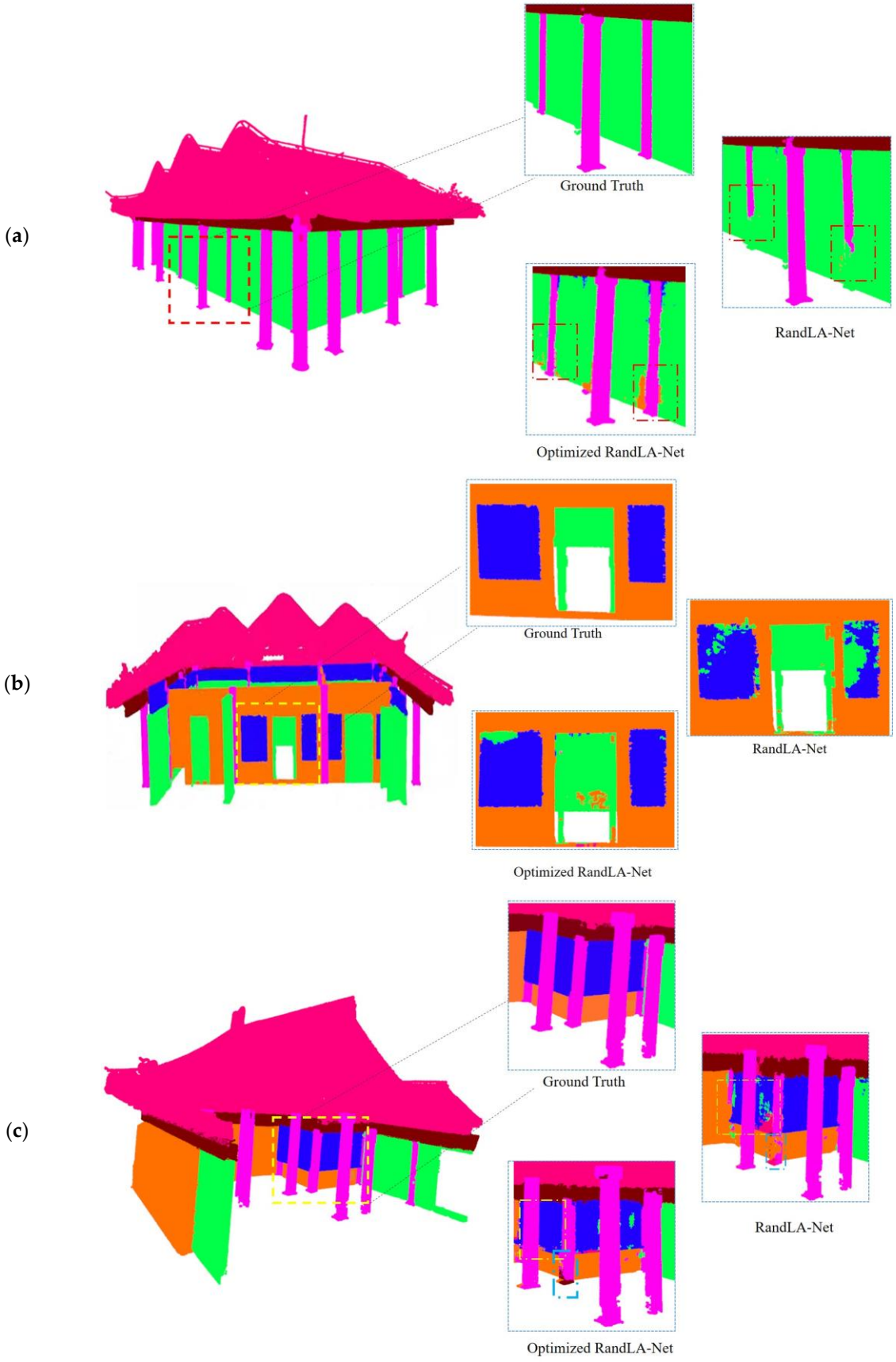

Figure 12. Comparison experiment visualization diagram.

4.2.2. Ablation Experiment of the CFE Module

We conducted an ablation study of the CFE module in area 4, and Table 5 displays the experimental outcomes. Firstly, all modules were removed, and then only the CFE module was added. The improvement from the first line to the second line shows that this module can fully consider the semantic gap between neighboring features and strengthen the connection between contexts. From the third to fourth lines, it can be seen that the effect of focusing only on partial features or global features is not ideal, focusing only on partial features will miss the semantic connection between neighbors, and focusing only on global features will lose a large number of detailed features, resulting in insufficient global feature mapping, and it is impossible to carry out more delicate semantic segmentation.

Table 5. Ablation experiment of the CFE module.

	mIoU (%)
(1) None	73.0
(2) CFE	75.3
(3) Only global feature	63.2
(4) Only partial feature	73.5

4.2.3. Ablation experiment of the EIC module

The EIC module introduces the features of each point to make better label predictions through the information transmission between each node, further improving the capacity to extract edge features to make the boundary segmentation clearer. We conducted ablation experiments in area 4, and Table 6 displays the experimental outcomes.

Table 6. Experimental results of EIC module for ablation.

	mIoU (%)
(1) None	73.0
(2) Only EIC	75.9
(3) Replace edge features with neighbor features	69.3
(4) Replace with average pooling	74.0

The table shows the quantitative results after adding the EIC module, replacing the edge feature with the neighbor feature, and replacing the average pooling module. As shown in Figure 12 above, without the addition of an edge classifier, the network cannot accurately identify the boundaries of objects that are easily misjudged, resulting in unclear edge segmentation of historical building components. The results in Table 6 show that the proposed EIC model has better prediction performance and more accurate classification. This indicates that the EIC module can significantly improve the accuracy of semantic segmentation of historical building components and make the segmentation boundary smoother.

5. Conclusions

This paper proposes an efficient MSFA-Net model to solve the issue of semantic segmentation of efficient architectural scene components. Three modules make up the model. The first module is made up of a bidirectional adaptive pooling block and a multiscale attention aggregation block that employs multilevel and different scale feature information to enhance the network's capacity to understand the topological relationship of nearby points and minimize redundant data. The second module, called the contextual feature enhancement module, combines local–global characteristics from the encoder and the decoder to enhance the relationship between the model contexts. As the third module, the edge interactive classifier further strengthens the extraction of edge features based on the original so that it can segment the edge of the object more smoothly.

Although this article has validated the superiority of the proposed model on both the public dataset S3DIS and the self-curated historical building dataset, there are still some

issues that need to be overcome. Firstly, research needs to continue to enrich the diversity of the dataset. Many types of historical buildings exist in the Chinese nation, and in the future, representative historical buildings from different periods need to be collected to enrich the types of components in the dataset and enhance the universality of the model. Secondly, due to the varying density of point clouds, the segmentation effect of building components with small data volumes and incomplete geometric information is poor. Therefore, future research will set constraint functions for different wooden components to further refine the research. Finally, with the continuous updating and development of collection instruments, the density and quality of point cloud data will continue to improve, and large-scale point cloud data will become more common, with higher annotation costs. Under this requirement, how to reduce model complexity and annotation costs will become the focus of future research.

Author Contributions: R.Z.: conceptualization, methodology, software, writing—review and editing. Y.X.: data curation, methodology, software, writing—original draft. J.W.: methodology, writing—review and editing, funding acquisition. D.S.: investigation, validation, writing—review and editing, funding acquisition. J.Z.: methodology, supervision, writing—review and editing. L.P.: supervision, writing—review and editing, Funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Natural Science Foundation of China (Grant No. 42171416); Beijing Municipal Natural Science Foundation (Grant No. 8222011).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to confidentiality restrictions.

Acknowledgments: The authors thank the managing editor and anonymous reviewers for their constructive comments.

Conflicts of Interest: Author Daixue Song, employed by Beijing Urban Construction Design & Development Group Co., Ltd., declares that he has no known competing financial interests or personal relationships that could be perceived as influencing the work reported in this paper. The remaining authors also declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Li, R.; Luo, T.; Zha, H. 3D Digitization and Its Applications in Cultural Heritage. In Proceedings of the Euro-Mediterranean Conference, Lemesos, Cyprus, 8–13 November 2010; pp. 381–388.
2. Ji, A.; Chew, A.W.Z.; Xue, X.; Zhang, L. An encoder-decoder deep learning method for multi-class object segmentation from 3D tunnel point clouds. *Autom. Constr.* **2022**, *137*, 104187. [CrossRef]
3. Xie, Y.; Tian, J.; Zhu, X.X. Linking points with labels in 3D: A review of point cloud semantic segmentation. *IEEE Geosci. Remote Sens. Mag.* **2020**, *8*, 38–59. [CrossRef]
4. Cheng, S.; Chen, X.; He, X.; Liu, Z.; Bai, X. Pra-net: Point relation-aware network for 3d point cloud analysis. *IEEE Trans. ImageProcess.* **2021**, *30*, 4436–4448. [CrossRef] [PubMed]
5. Chen, Y.; Liu, X.; Xiao, Y.; Zhao, Q.; Wan, S. Three-Dimensional Urban Land Cover Classification by Prior-Level Fusion of LiDAR Point Cloud and Optical Imagery. *Remote Sens.* **2021**, *13*, 4928. [CrossRef]
6. Pérez-Sinticala, C.; Janvier, R.; Brunetaud, X.; Treuillet, S.; Aguilar, R.; Castañeda, B. Evaluation of Primitive Extraction Methods from Point Clouds of Cultural Heritage Buildings. In *Structural Analysis of Historical Constructions*; RILEM Bookseries; Springer: Cham, Switzerland, 2019; pp. 2332–2341.
7. Kivilcim, C.Ö.; Duran, Z. Parametric Architectural Elements from Point Clouds for HBIM Applications. *Int. J. Environ. Geoinformatics* **2021**, *8*, 144–149. [CrossRef]
8. Cheng, M.; Hui, L.; Xie, J.; Yang, J. Sspc-net: Semi-supervised semantic 3d point cloud segmentation network. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 1140–1147. [CrossRef]
9. Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; Bennamoun, M. Deep learning for 3d point clouds: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 4338–4364. [CrossRef] [PubMed]
10. Le, T.; Duan, Y. Pointgrid: A Deep Network for 3D shape understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 8–23 June 2018; pp. 9204–9214.
11. Meng, H.Y.; Gao, L.; Lai, Y.K.; Manocha, D. Vv-net: Voxel vae net with group convolutions for point cloud segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8500–8508.

12. Milioto, A.; Vizzo, I.; Behley, J.; Stachniss, C. RangeNet ++: Fast and Accurate LiDAR Semantic Segmentation. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 4213–4220.
13. Lyu, Y.; Huang, X.; Zhang, Z. Learning to segment 3d point clouds in 2d image space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12255–12264.
14. Triess, L.T.; Peter, D.; Rist, C.B.; Zöllner, J.M. Scan-based Semantic Segmentation of LiDAR Point Clouds: An Experimental Study. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; pp. 1116–1121.
15. Chen, Y.; Liu, G.; Xu, Y.; Pan, P.; Xing, Y. PointNet++ Network Architecture with Individual Point Level and Global Features on Centroid for ALS Point Cloud Classification. *Remote Sens.* **2021**, *13*, 472. [CrossRef]
16. Qian, G.; Hammoud, H.; Li, G.; Thabet, A.; Ghanem, B. ASSANet: An Anisotropic Separable Set Abstraction for Efficient Point Cloud Representation Learning. *Neural Inf. Process. Syst.* **2021**, *34*, 28119–28130.
17. Qian, G.; Li, Y.; Peng, H.; Mai, J.; Hammoud, H.; Elhoseiny, M.; Ghanem, B. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 23192–23204.
18. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
19. 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, Seattle, WA, USA, 13–19 June 2020; pp. 11108–11117.
20. Fan, S.; Dong, Q.; Zhu, F.; Lv, Y.; Ye, P.; Wang, F.Y. SCF-Net: Learning Spatial Contextual Features for Large-Scale Point Cloud Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 14499–14508. [CrossRef]
21. Zeng, Z.; Xu, Y.; Xie, Z.; Tang, W.; Wan, J.; Wu, W. LACV-Net: Semantic Segmentation of Large-Scale Point Cloud Scene via Local Adaptive and Comprehensive VLAD. *arXiv* **2022**, arXiv:2210.05870.
22. Mao, Y.; Sun, X.; Chen, K.; Diao, W.; Guo, Z.; Lu, X.; Fu, K. Semantic segmentation for point cloud scenes via dilated graph feature aggregation and pyramid decoders. *arXiv* **2022**, arXiv:2204.04944.
23. Xue, Y.; Zhang, R.; Wang, J.; Zhao, J.; Pang, L. EEI-NET: EDGE-ENHANCED INTERPOLATION NETWORK FOR SEMANTIC SEGMENTATION OF HISTORICAL BUILDING POINT CLOUDS. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2023**, *10*, 239–245. [CrossRef]
24. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
25. Armeni, I.; Sax, S.; Zamir, A.R.; Savarese, S. Joint 2D-3D-semantic data for indoor scene understanding. *arXiv* **2017**, arXiv:1702.01105.
26. Shuai, H.; Xu, X.; Liu, Q. Backward Attentive Fusing Network With Local Aggregation Classifier for 3D Point Cloud Semantic Segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 4973–4984. [CrossRef] [PubMed]
27. Su, Y.; Liu, W.; Yuan, Z.; Cheng, M.; Zhang, Z.; Shen, X.; Wang, C. DLA-Net: Learning dual local attention features for semantic segmentation of large-scale building facade point clouds. *Pattern Recognit.* **2022**, *123*, 108372. [CrossRef]
28. Landrieu, L.; Simonovsky, M. Large-scale point cloud semantic segmentation with superpoint graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4558–4567.
29. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. Pointcnn: Convolution on x-transformed points. *arXiv* **2018**, arXiv:1801.07791.
30. Zhao, H.; Jiang, L.; Fu, C.W.; Jia, J. Pointweb: Enhancing local neighborhood features for point cloud processing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5565–5573.
31. Zhang, Z.; Hua, B.S.; Yeung, S.K. Shellnet: Efficient point cloud convolutional neural networks using concentric shells statistics. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1607–1616.