remote sensing

MDPI

Technical Note

Multi-Feature Aggregation for Semantic Segmentation of an Urban Scene Point Cloud

Jiaqing Chen [1,2], Yindi Zhao [1,2,*], Congtang Meng [2] and Yang Liu [2]

[1] Key Laboratory of Degraded and Unused Land Consolidation Engineering, Ministry of Natural Resources, Xi'an 710075, China

[2] School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China

* Correspondence: zhaoyd@cumt.edu.cn

Abstract: With the rapid development of cities, semantic segmentation of urban scenes, as an important and effective imaging method, can accurately obtain the distribution information of typical urban ground features, reflecting the development scale and the level of greenery in the cities. There are some challenging problems in the semantic segmentation of point clouds in urban scenes, including different scales, imbalanced class distribution, and missing data caused by occlusion. Based on the point cloud semantic segmentation network RandLA-Net, we propose the semantic segmentation networks RandLA-Net++ and RandLA-Net3+. The RandLA-Net++ network is a deep fusion of the shallow and deep features of the point clouds, and a series of nested dense skip connections is used between the encoder and decoder. RandLA-Net3+ is based on the multi-scale connection between the encoder and decoder; it also connects internally within the decoder to capture fine-grained details and coarse-grained semantic information at a full scale. We also propose incorporating dilated convolution to increase the receptive field and compare the improvement effect of different loss functions on sample class imbalance. After verification and analysis of our labeled urban scene LiDAR point cloud dataset—called NJSeg-3D—the mIoU of the RandLA-Net++ and RandLA-Net3+ networks is 3.4% and 3.2% higher, respectively, than the benchmark network RandLA-Net.

Keywords: point cloud semantic segmentation; multi-feature aggregation; RandLA-Net; deep learning

check for updates

Citation: Chen, J.; Zhao, Y.; Meng, C.; Liu, Y. Multi-Feature Aggregation for Semantic Segmentation of an Urban Scene Point Cloud. *Remote Sens.* **2022**, *14*, 5134. https://doi.org/10.3390/rs14205134

Academic Editor: Mostafa Arastounia

Received: 1 September 2022
Accepted: 12 October 2022
Published: 14 October 2022

1. Introduction

1.1. Background

The rational spatial layout and land use mode of a city can stimulate the potential efficiency of land use to the greatest extent, bring the maximum economic and social benefits, forming a competitive and sustainable land space pattern. By semantic segmentation of typical urban ground features, the spatial layout can be obtained accurately and quickly, thus playing an important and significant role in urban development planning, smart city construction, land management, geographic database update, and other practical fields [1].

As one of the most important research techniques in computer vision, semantic segmentation was first proposed in the 1970s to classify each pixel or point in a scene into several regions using specific semantic categories. Currently, semantic segmentation technology based on two-dimensional imaging has been maturely developed, e.g., U-Net [2], DeepLab [3], SegNet [4], etc., continuously improve the accuracy of image semantic segmentation. For a long time, two-dimensional spatial data represented by maps and images were far from meeting the needs of various applications [5]. In recent years, the innovation and development of earth-to-earth observation technologies, such as intelligent small satellites, low-altitude UAVs, and ground mobile three-dimensional scanning and measurement, have rapidly improved the perception capabilities for all space and time domains [6].

A point cloud is a set of data points in space. The points may represent a 3D shape or object. With the maturity of depth cameras, 3D laser scanners, LiDAR, and other technologies, the number of 3D point cloud datasets is also increasing, e.g., ShapeNet [7], ModelNet [8], ScanNet [9], Semantic3D [10], SensatUrban [11], Toronto-3D [12], etc. These technologies, especially the ability to rapidly extract 3D geographic information [13], make the semantic segmentation of large-scale 3D point clouds possible, which promotes the wide application of 3D point cloud segmentation in automatic driving, intelligent robots, smart cities, remote sensing mapping, and other fields. In this work, we focused on the semantic segmentation of 3D point clouds in urban scenes based on deep learning models, which were improved by multi-feature fusion, and obtained robust results.

1.2. Related Work

With the emergence of deep learning technology, great improvements have been achieved in the field of point cloud semantic segmentation [14]. In recent years, researchers have proposed a large number of segmentation models based on deep learning to deal with point clouds. According to whether point clouds requires structured preprocessing, deep learning-based point cloud semantic segmentation methods are mainly divided into indirect point cloud semantic segmentation and direct point cloud semantic segmentation [15]. The indirect semantic segmentation method is used to convert original point cloud data into conventional 3D voxel mesh or project-based materials, and to extract features from 3D point cloud data indirectly through data transformation, so as to achieve the purpose of the semantic segmentation. The direct semantic segmentation method is used to extract feature information directly from the point cloud data. In the case where there is no transformation into voxels and multi-views, the architecture retains the inherent information in the original points to predict the point-level semantics.

Specifically, indirect semantic segmentation methods can be divided into project-based and 3D voxel-based methods. The projection-based method is used to project the 3D point cloud to a 2D image and then to use a convolutional neural network for semantic segmentation, such as the multi-view convolutional neural network (MVCNN) proposed by Su et al. [16]. The method considers projecting from many different angles to obtain a multi-view 2D image. However, this method will lead to the loss of spatial geometric information due to projection, which leads to a reduction in segmentation accuracy. Methods based on voxelization convert unstructured point clouds into regular 3D grids and then use neural networks to learn their features to realize semantic segmentation of the point cloud. For example, Maturana et al. [17] proposed the VoxNet model and used the volume method to convert unstructured geometric data into conventional 3D grids that can perform standard CNN operations. The 3D-CNN is then used to predict labels directly from the 3D grid. Due to the high space complexity of the voxelization algorithm, this method requires a large overhead in the process of storage and operation.

From the above literature analysis, it can be seen that the point cloud semantic segmentation model based on the indirect method can solve the issue that CNN cannot be directly applied to point cloud operations, but there are problems with this model, such as information loss and large memory usage, which must to be further improved. The method of directly processing unstructured point clouds fundamentally avoids the problem of spatial geometric information loss; thus, it is an important research direction for current point cloud semantic segmentation. This method is divided into four main categories: the point-wise MLP method, the point convolution method, graph-based methods and RNN-based methods.

Pointwise MLP methods use a shared MLP as the basic unit, and the PointNet [18] network is a pioneer in this class of methods, establishing the direct application of deep learning to point cloud semantic segmentation while addressing permutation invariance (i.e., the order of the input point cloud should not change the object category represented by the point cloud) and geometric transformation invariance (i.e., the point cloud should still represent the same object after translation and rotation). PointNet can be directly applied to point cloud processing, and it has been successfully applied to indoor point cloud data segmentation, but because it does not consider the contextual relationship between the current point and local neighborhood points, the segmentation effect is not ideal when processing point clouds of larger scenes. In the same year, Qi et al. released an upgraded version of PointNet, PointNet++ [19]. The PointNet++ network consists of a sampling layer, a grouping layer, and a PointNet layer. First, the farthest point sampling algorithm (FPS) is used to select several points from the input point cloud as the centroid of the local area, and then the local area grouping module is added to construct local regions; finally, PointNet is used to recursively extract the local features. Although the upgraded version of PointNet++ effectively solves the problem of local feature extraction and improves the segmentation accuracy, it still processes each point independently, without considering information such as distance and direction between points. Hu et al. comprehensively studied the problems of the PointNet and PointNet++ models, fully considering the spatial relationship between points in the point cloud, and proposed a large-scene point cloud semantic segmentation network RandLA-Net [20] based on random sampling. RandLA-Net is a lightweight network for large-scale point cloud processing which uses the random point sampling method instead of the farthest point sampling method, and achieves significant improvement in storage and computation through a local feature aggregation module to capture and retain local geometric features.

Due to the disorder of the point cloud data, the arrangement order of the input point cloud data varies significantly, making it difficult to directly apply the convolution operation to the point cloud data. To further address this problem and take advantage of standard CNN operations, PointCNN [21] attempts to learn the χ-transform convolution operator to transform the unordered point cloud into the corresponding canonical order, after which the typical CNN architecture is used to extract local features. Similarly, in the process of solving for the lack of spatial convolution, Thomas et al. [22] proposed kernel point convolution (KPConv), which provides a deformable convolution operator. By applying the weight of the nearest kernel point in the neighborhood, each local neighborhood is convoluted. The convolution weight of KPConv is determined by the Euclidean distance to the kernel point, and the number of kernel points is not fixed, so KPConv is more flexible than fixed grid convolution.

In addition, in regards to graph convolution-based methods, Wang et al. [23] proposed a local spectral graph convolution, which constructs a local graph from the neighborhood of a point and uses spectral graph convolution to combine a new graph pooling strategy to learn the relative layout and features of adjacent points. Wang et al. [24] improved the previous method and proposed a dynamic graph convolutional neural network (DGCNN). DGCNN constructs a local neighborhood map and uses edge convolution operation to extract the feature and the edge vector of the center point and K-nearest neighbor (KNN) points to obtain the local features of the point cloud.

A recurrent neural network (RNN) [25] is another mainstream model in deep learning. RNN cannot only learn the information of the current moment, but can also rely on the previous sequence information, which is conducive to modeling global content and saving history data to facilitate the use of contextual information. Paper [26] fused the 3D convolutional neural network, deep Q-network (DQN), and residual recurrent neural network (RNN), and proposed the 3DCNN-DQNRNN for the semantics of large-scale point clouds parsing. Paper [27] considered improving 3D point cloud segmentation results by exploiting the initial segmentation scores of neighboring points, and proposed an attention-based score refinement (ASR) module in which the weights are calculated from the initial segmentation scores of the points, and the scores of each point and its neighbors are combined based on the calculated weights to optimize the scores. This module can be easily integrated into existing deep networks to improve the final segmentation effect.

From the above literature analysis, we can see that the current point cloud semantic segmentation algorithm based on deep learning has achieved some successes, but there is still space for improvement in the utilization of feature information. Therefore, we have mainly focused on two aspects for improved. On the one hand, the shallow features and deep features of the point cloud are fused to meet the needs of different network depths and to improve the overall segmentation accuracy. On the other hand, full-scale skip connections combine high-level semantics with low-level semantics from feature maps at different scales. Furthermore, dilated convolutions are also introduced into our model to enlarge the receptive field. The main contributions of this paper include the following:

(1) To bridge the semantic gap between point cloud features in the encoder and decoder, a densely connected module is proposed to replace traditional long skip connections. Multi-scale aggregation is introduced to fuse point cloud features of different scales from the decoder and the encoder. In the dilated residual module, we add the dilated convolutions to further increase the receptive field.

(2) Our RandLA-Net++ and RandLA-Net3+ significantly outperformed the state-of-art methods using the SensatUrban, Toronto-3D, and NJSeg-3D datasets, achieving mean Intersection over Union (mIoU) scores of at least 3% and 2%, respectively.

2. Study Area and Materials

We take the Pukou District of Nanjing City, which covers 30.1 km^2, as the research area. The point cloud data and UAV images were acquired by airborne LiDAR and UAV, respectively. The collected data were fused by MicroStation's Terrascan software (Bentley Systems Incorporated, Exton, PA, USA). The processed point cloud density is 16 points/m^2, the intensity range is 0–65,535, and the attribute information is the XYZ coordinates, the intensity, and the RGB spectral information. Figure 1a represents the unlabeled original point clouds; Figure 1b represents the labeled point clouds.

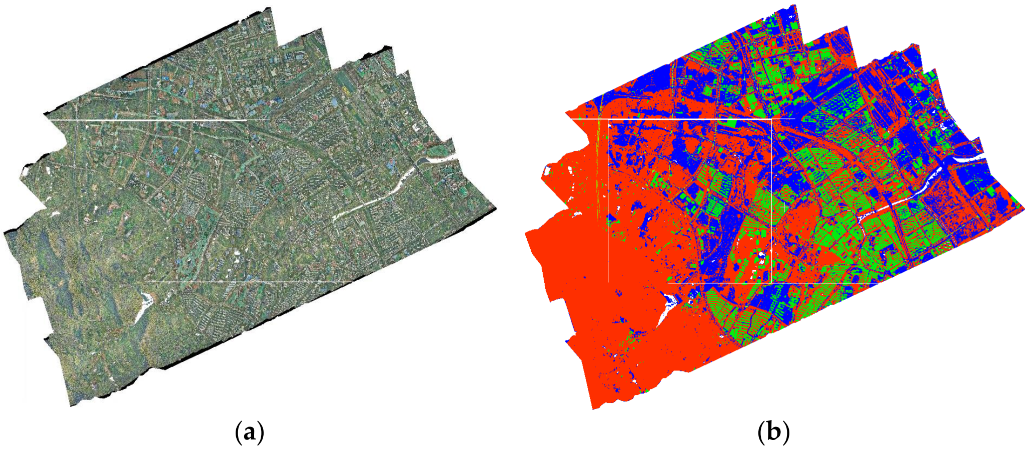

Figure 1. Original point cloud and labeled point cloud: (**a**) point cloud before labeling; (**b**) labeled point cloud.

We refer to the Semantic3d format standard of the point cloud semantic segmentation dataset, and label the fused point cloud data with buildings, vegetation, roads, and four other types of objects and obtain a point cloud semantic segmentation dataset named NJSeg-3D; some of the labeled point cloud data are shown in Figure 2. There are 45 tiles in our NJSeg-3D dataset, among which are 27 tiles for training, 6 for validation, and 12 for testing. Each tile covers an area of approximately 0.8 km by 0.8 km. The RandLA-Net/RandLA-Net++/RandLA-Net3+ model is trained on the training tiles, tuned the hyper-parameters on the validation tiles, and evaluated on the test tiles.

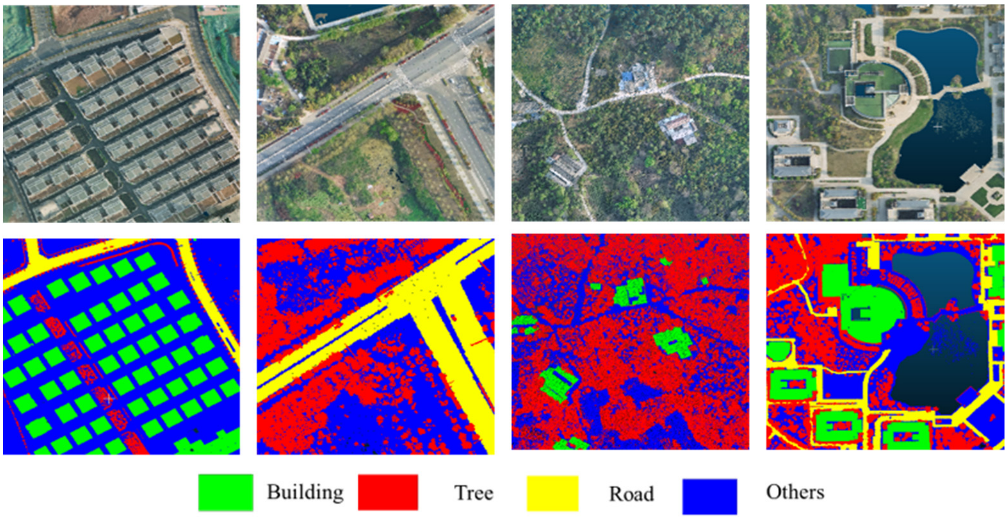

Figure 2. Examples of the NJSeg-3D dataset. Different semantic classes are labeled using different colors, where green point clouds represent buildings, red point clouds represent trees, yellow point clouds represent roads, and blue point clouds represent other objects.

3. Methodology

Point clouds in urban scenes contain problems such as different scales, imbalanced class distribution, data loss due to occlusion, as well as others (as shown in Figure 3). Based on the RandLA-Net point cloud semantic segmentation network, inspired by the UNet series of two-dimensional image semantic segmentation networks [2,28,29], we fuse point cloud features of different depths and scales to enhance the feature representation. At the same time, we add dilated convolutions with different dilation rates to increase its receptive field, coping with the imbalanced class distribution problem by using a more complex loss function.

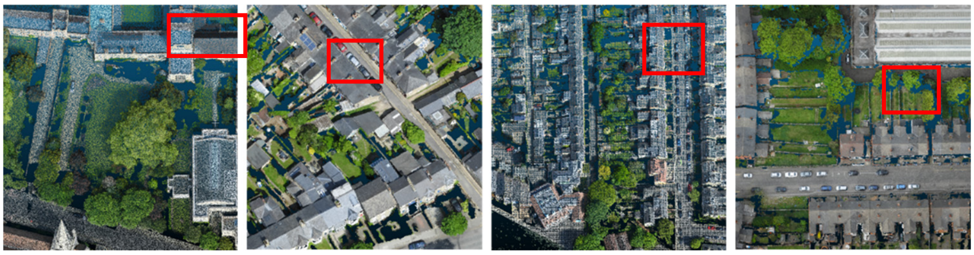

Figure 3. Part of the point cloud data (with missing data) in the SensatUrban dataset. The blue dots correspond to missing information, such as that shown in the red boxes.

3.1. RandLA-Net Network

RandLA-Net is an end-to-end direct point cloud semantic segmentation network that adopts the idea of encoding and decoding with a skipping connection and takes multi-layer perceptron (MLP) as the basic unit. It is an efficient neural network based on the random sampling (RS) principle and the local feature aggregation (LFA) module. When processing large scene point clouds, based on the multi-layer perceptron module, no extra preprocessing or post-processing steps are required, which reduces the memory consumption of 3D point cloud semantic segmentation so that it can be quickly segmented in large-scale point clouds. At the encoding end, the point cloud enriches and learns the features of the points at each layer through the local feature aggregation algorithm and uses random sampling to reduce the scale of the point cloud. The decoding side uses the linear interpolation method of each point and KNN to obtain the nearest point for up-sampling, superpositions the features of the encoding side through skip connection, and then inputs the shared MLP for feature dimensionality reduction. The last three fully connected layers and the dropout layer are used to predict the category of each point. The network structure is shown in Figure 4.

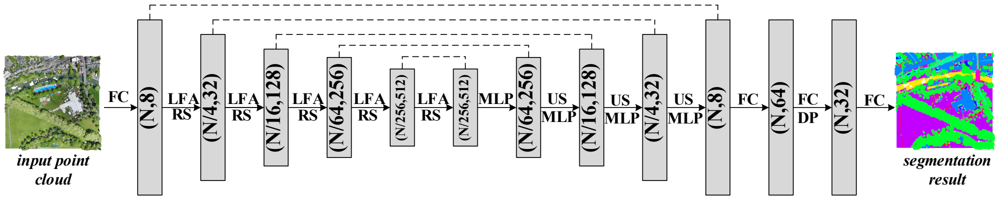

Figure 4. The structure of RandLA-Net. The numbers in the grey block show the number of channels of the feature and the number of points of the input, respectively. FC, LFA, RS, MLP, US, DP, and N represent the fully connected layer, local feature aggregation, random sampling, multi-layer perceptrons, up-sampling, dropout, and the number of input points (for example, N = 65,536), respectively.

The local feature aggregation (LFA) algorithm is the core of the RandLA-Net structure. It is mainly composed of three modules, including local spatial encoding (LocSE), attention pooling, and the dilation residual block. Among these, the LocSE uses MLP to encode the position information of each point and the nearest neighbor points of the input point cloud. The information used includes the position of the point p_i, the position of the adjacent point p_{ik}, the relative position relationship between the point and the adjacent point $p_i - p_{ik}$, and their Euclidean distance $\|p_i - p_{ik}\|$. An attention pooling module is used to aggregate the local features of the points. Most of the existing works use maximum or average pooling to force the integration of adjacent features, resulting in the loss of a large amount of information; therefore, important local features are automatically learned by using the attention mechanism. In order to deal with large-scale point clouds, RandLA-Net adopts the random sampling method to down-sample the point cloud at the encoding end, since it has the characteristics of low time and space complexity, high efficiency, and a small memory footprint. However, random sampling may cause the loss of key point information. Therefore, the dilated residual block module is used to stack local spatial encoding, attention pooling, and multi-layer perceptrons to increase the receptive field of each point and reduce the impact of the random sampling loss of key point information.

3.2. Multi-Feature Aggregation

The RandLA-Net network follows the method of encoding and decoding, extracts the features of the point cloud at the encoding end, and restores the extracted features to the information of the required size at the decoding end to realize the semantic segmentation of the point cloud. It has shown excellent performance in large-scale semantic segmentation, but there is still room for improvement in both the selection of the depth that optimizes the model performance [28] and the model semantic differences caused by the different expressive capabilities of features at different depths [30]. In real scenarios, the complexity and data volume of point cloud data in different scenarios will vary, and the fixed depth model of the RandLA-Net network limits its choice of the best depth of the model; thus, directly connecting the encoder and decoder in the RandLA-Net network will cause a large semantic difference. For example, as shown in Figure 5, the output of node $X^{1,0}$ in RandLA-Net has undergone slight transformations (only few local feature aggregation and MLP operations), while the output of $X^{1,4}$ has undergone nearly all possible transformations (four down-sampling and three up-sampling stages). Therefore, there is a significant gap between the representation capabilities of $X^{1,0}$ and $X^{1,4}$. Therefore, simply connecting the outputs of $X^{1,0}$ and $X^{1,4}$ is not the best solution. At the same time, there are additional problems, including different scales and imbalanced class distribution, inherent in the point cloud of urban scenes. In addition, the point cloud data is down-sampled many times, which will forfeit many details. It is necessary to retain as many of the geometric details of the input point cloud as possible. The feature learning ability is improved through multi-feature fusion, which provides help for target extraction and category distinction. Based on this, two point cloud semantic segmentation networks, RandLA-Net++ and RandLA-Net3+, are proposed according to the feature fusion of different depths and scales of point clouds. The RandLA-Net++ point cloud semantic segmentation network realizes the fusion of both shallow and deep features of point clouds. In this paper, shallow features are low-level features such as edges, shapes, and colors, while deep features are high-level features that measure object characterization, classification, and scene parsing. The redesigned skip connections change the dense connections of the encoder and decoder to bridge the semantic gap between the feature maps of the encoder and decoder; the RandLA-Net3+ point cloud semantic segmentation network utilizes the multi-scale features of point clouds and uses full-scale skip connections to combine high- and low-level semantics from feature maps of different scales to realize the fusion of features of different scales.

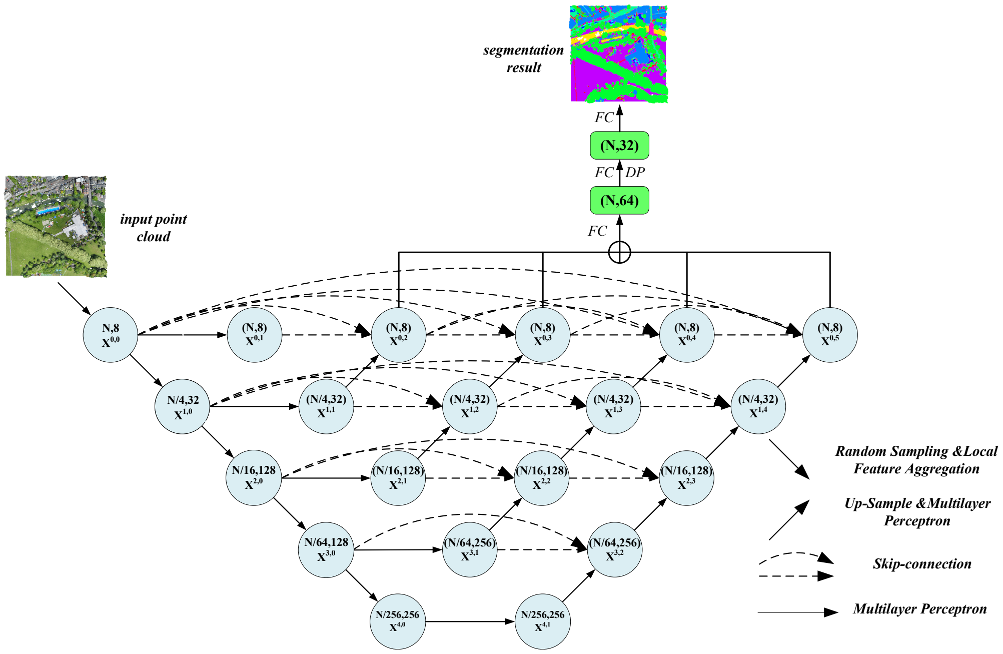

Figure 5. The structure of RandLA-Net++. The numbers in the circles show the number of channels of the feature and the number of input points, respectively.

3.2.1. RandLA-Net++ Network

The RandLA-Net++ network consists of RandLA-Net networks of different depths whose decoders are densely connected at the same resolution through redesigned skip connections. Compared to RandLA-Net, RandLA-Net++ has the following advantages. First, it embeds RandLA-Nets of different depths. These RandLA-Nets share an encoder and are densely connected at the decoding end to meet the needs of different network depths and improve the overall segmentation accuracy. Second, it is not limited by unnecessary skip connections; that is, only feature maps of the same scale of the encoder and decoder can be fused. The introduced redesigned skip connections present feature maps of different scales on the decoder node, allowing the aggregation layer to decide how to fuse the various feature maps carried by skip connections with the decoder features, bridging the semantic gap between the feature maps of the encoder and decoder. The skip link is represented as follows: let $x^{i,j}$ denote the output of node $X^{i,j}$, where i indexes the down-sampling layers along the encoder, and j indexes the multilayer perceptron layers of dense blocks along the skip path. The stack of feature maps denoted by $x^{i,j}$ is computed as:

$$x^{i,j} = \begin{cases} D(x^{i-1,j}), & j = 0 \\ mlp(x^{i,j-1}), & j = 1 \\ mlp([[x^{i,n}]_{\mathrm{n=0}}^{j-1}, U(x^{i+1,j-1})]), & j > 1 \end{cases} \quad (1)$$

where the function *mlp*(·) is a multilayer perceptron operation, *D*(·) denotes an down-sampling layer, *U*(·) denotes an up-sampling layer, and [·] denotes a connection layer.

The redesigned skip connection is implemented in RandLA-Net++ by densely connecting the decoders of RandLA-Net with the same resolution, which is equivalent to fusing the output of the previous layer with the corresponding up-sampling of the lower layer; thus, a semantically more similar feature map is obtained, which is beneficial for the optimization of the model. Dense skip connections ensure that the previously accumulated feature maps are utilized, enabling multiple semantic levels to generate full-resolution feature maps, thereby improving the semantic segmentation accuracy of 3D point clouds.

The network structure is shown in Figure 5. The random sampling and local aggregation modules are combined together. The input point cloud is continuously down-sampled in RandLA-Net++ to save computing resources and memory overhead. The features of each layer at the encoding end are processed by MLP. Then, nearest neighbor interpolation is used to achieve up-sampling, and at the decoding end, the features of the corresponding depth are densely connected through the dense connection operation, where the up-sampling operation is used to select a more efficient nearest neighbor interpolation. Finally, two fully connected layers, one dropout layer, and one fully connected layer are connected to output the corresponding point cloud object category.

3.2.2. RandLA-Net3+ Network

In the semantic segmentation of point clouds in urban scenes, there is a large difference in the scale of ground objects. In the feature extraction of point cloud semantic segmentation, features of different scales play different roles. The low-level features obtain rich spatial information, while the high-level features contain semantic information. However, due to the multiple down-sampling and nearest-neighbor interpolation operations brought about by its encoding-decoding structure, some information may be lost. To take full advantage of multi-scale features, we redesign the connections between the encoder and decoder and the internal connections between the decoders to capture full-scale fine-grained details and coarse-grained semantic information. The skip link is represented as follows: let $\boldsymbol{x^{i,j}}$ denote the output of node $X^{\boldsymbol{i,j}}$, i represents the i^{th} down-sampling layer along the encoding direction, and j represents the j^{th} up-sampling layer along the decoding direction; then, the calculation formula of the feature map is as follows:

$$x^{i,j} = \begin{cases} D(x^{i-1,j}), & j = 0 \\ mlp(x^{i,j-1}), & j = 1 \\ mlp([D[x^{k,0}]_{\mathrm{k=0}}^{i-1}, x^{i,0}, U([x^{\mathrm{k},n}]_{k=i+1}^{4})_{n=1}^{j-1}]), & j > 1 \end{cases} \quad (2)$$

Among these, where the function *mlp*(·) is a multilayer perceptron operation, *D*(·) denotes an down-sampling layer, *U*(·) denotes an up-sampling layer, and [·] denotes a connection layer.

The network structure is shown in Figure 6. A full-size skip connection is used between the encoder and the decoder, and the multi-scale features are internally connected at the decoding end, thereby making up for the ordinary connected RandLA-Net and the nested and densely connected RandLA-Net. We propose RandLA-Net3+ network, in which we redesign the interconnections between the encoder and the decoder and add the intraconnections among the decoder layers, which enable to capture fine-grained details and coarse-grained semantic information from the full-scale feature maps. The specific operation is to down-sample the features of each layer at the encoding end through maximum pooling to obtain the scale features corresponding to the decoding end, followed by fusing; at the decoding end, the output features of each layer are interpolated to the corresponding scale, and nearest neighbor interpolation is used to obtain features of corresponding scale. Then, the splicing operation is performed with the features at the encoding end, thereby obtaining the point cloud features with multi-scale fusion. For the fusion operation between different scales, we use the concatenation operation. Finally, two fully connected layers, one dropout layer, and one fully connected layer are connected to output the corresponding point cloud object category.

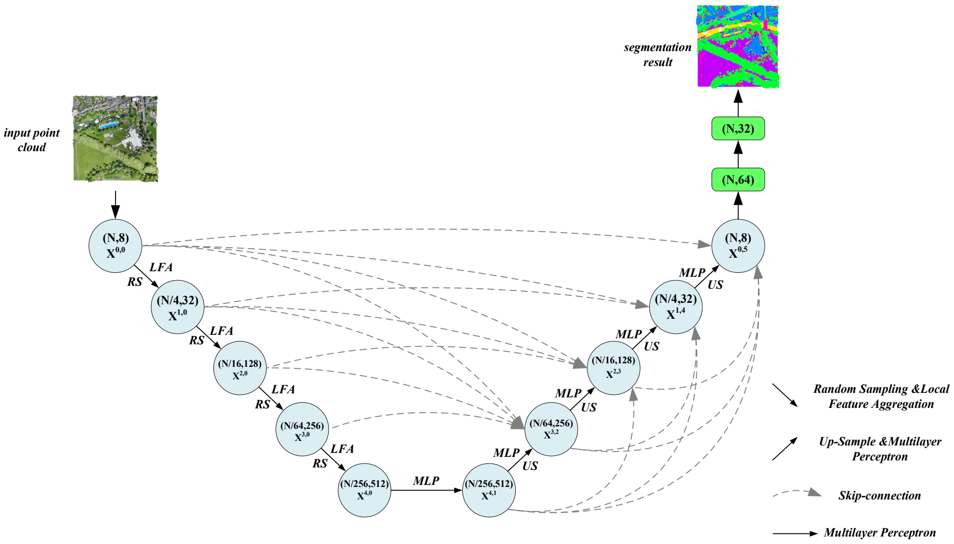

Figure 6. The structure of RandLA-Net3+. The numbers in and below the circles show the number of channels of the feature and the number of input points, respectively. RS, US, and MLP are used for random sampling, up-sampling, and multilayer perceptrons, respectively.

3.3. Dilated Convolution

In order to further reduce the loss of spatial and semantic feature information from the input point cloud, a fusion module is designed to enhance the receptive field. Dilated convolution is added to the dilated residual block module. Dilated convolution is a special structure of the convolution operation which can expand the receptive field of the convolution kernel without adding parameters. By setting different expansion rates, the multi-scale feature information of the target can be obtained, which is beneficial for the extraction of the target features. Convolutions with different dilation rates (r = 1, r = 5) are used to increase the receptive field and preserve more geometric and semantic features. Figure 7 is a structural diagram improved by adding dilated convolution based on the dilated residual module in the RandLA-Net model [20]. As shown in the red dashed box in Figure 7, by fusing the features obtained from the shared MLP and the features obtained using the dilated convolution, the large receptive field features can be obtained.

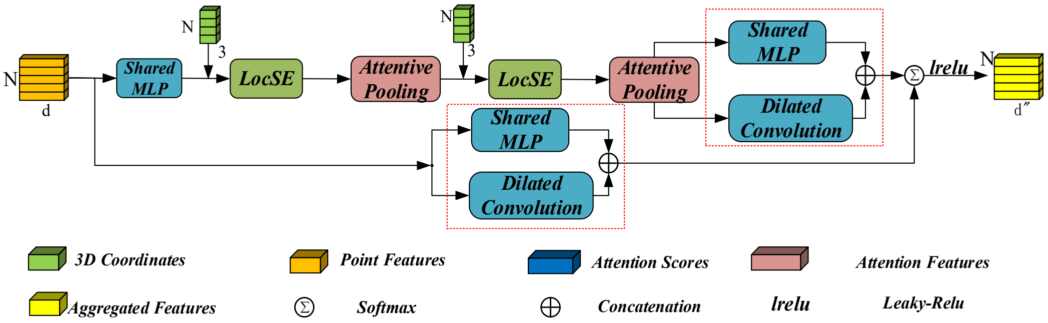

Figure 7. Receptive field enhancement fusion module.

3.4. Loss Function

To alleviate the class imbalance of urban scene point clouds, some loss functions are introduced, such as generalized dice loss [31], focal loss [32], weighted cross-entropy loss [33], and Lovász-Softmax loss [33]. In this paper, we use weighted cross-entropy loss, Lovász-Softmax loss, and the combined loss of weighted cross-entropy and Lovász-Softmax loss.

The weighted cross-entropy loss function [33] is defined as follows:

$$\mathrm{L_{wce}}(y, \hat{y}) = -\sum_i \alpha_i p(y_i) \log(p(\hat{y}_i)) \tag{3}$$

where $\alpha_i = 1/\sqrt{f_i}$, y_i and $\hat{y}_i$ represent the true and predicted class labels, and f_i represents the frequency of the i^{th} class, i.e., the number of points, which strengthens the network's response to classes with fewer occurrences in the dataset.

The Lovász-Softmax loss function [33] can be formulated as follows:

$$\mathrm{L_{ls}} = \frac{1}{|C|} \sum_{c \in C} \overline{\Delta_{J_c}}(m_\mathrm{i}(c)) \tag{4}$$

where

$$m_i(c) = \begin{cases} 1 - x_i(c) & if c = y_i(c) \\ x_i(c) & otherwise \end{cases} \tag{5}$$

where ∣C∣ represents the class number, $\overline{\Delta_{J_c}}$ defines the Lovász extension of the Jaccard index, x_i(c) $\in$ [0,1] and $y_i(\mathrm{c}) \in \{-1, 1\}$ hold the predicted probability and ground truth label of point cloud i for class c, respectively.

Finally, the combined loss function of RandLA-Net++ is a linear combination of the weighted cross-entropy loss and the Lovász-Softmax loss, that is, L = $\mathrm{L_{wce}}$ + $\mathrm{L_{ls}}$.

4. Results

4.1. Experiment Design

We evaluated the proposed method using three datasets—SensatUrban, Toronto-3D, and the NJSeg-3D dataset. Our experimental environment and parameter settings are as consistent as possible with the RandLA-Net model. Our experiments were carried out under the Ubuntu system, and the number of points input to the model is 65,536, the k value of the KNN is 16, the initial learning rate is 0.01, the momentum is 0.95, and the model iteration is 100.

4.2. Evaluation Metrics

For better comparison with the RandLA-Net baseline model, we keep the evaluation metrics consistent. For the SensatUrban and Toronto-3D datasets, we use the Overall Accuracy (OA) and mean Intersection over Union (mIoU). For the NJSeg-3D dataset, we provide other additional evaluation metrics, such as precision, recall, and F1 score.

The OA can judge the overall correct rate, which is defined as the percentage of correct prediction results in the overall population. The formula is:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \tag{6}$$

where TP, TN, FP, and FN are the true positive case, true negative case, false-positive case, and false-negative case, respectively.

The mean Intersection over Union (mIoU) is used to calculate the ratio of the intersection and union of all category prediction results and the true value:

$$\mathrm{mIoU} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{\mathrm{TP}}{\mathrm{FN} + \mathrm{FP} + \mathrm{TP}} \tag{7}$$

where *k* represents the category and *k* + 1 means that the background category has been added.

Precision is the percentage of the predicted correct results among the predicted results of all positive samples. The formula is:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

Recall is the percentage of true positives predicted to be correct; the formula is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

The F1 Score is a comprehensive evaluation index for evaluating precision and recall; the formula is:

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

4.3. Experimental Results and Analysis

This paper is mainly based on the deep learning method of RandLA-Net, and it proposes two different multi-feature fusion networks: RandLA-Net++ and RandLA-Net3+. We conduct experimental analysis on the urban scene data acquired by different devices, such as the photogrammetric point cloud SensatUrban dataset for large outdoor scenes, the urban scene point cloud Toronto-3D, and the LiDAR point cloud NJSeg-3D of urban scenes.

4.3.1. SensatUrban Dataset

The SensatUrban dataset [11] is a large-scale photogrammetric point cloud dataset of urban scenes, with a total of 13 object categories, nearly 7.6 square kilometers of scene, and close to 3 billion points with semantic annotation. This dataset is three times the size of the largest previously existing photogrammetric point cloud dataset. In addition to more common categories, such as roads and vegetation, city-level categories, such as railways, bridges, and rivers, are also included. We keep the same training, validation, and test datasets used by RandLA-Net. The SensatUrban dataset represents sub-sections of three large cities in the UK, i.e., Birmingham, Cambridge, and York. In particular, the point cloud of the Birmingham urban area is divided into 14 tiles. We then select 10 tiles for training, 2 for validation, and 2 for testing. Similarly, the Cambridge split has 29 tiles in total: 20 tiles for training, 5 for validation, and 4 for testing. Each tile is approximately 400 × 400 square meters.

For a fairer comparison of our network with the RandLA-Net network and to highlight the superiority of our model, we show that the improvement in the prediction accuracy of our network comes from the improvement of the model rather than from relying on the loss function. For the SensatUrban dataset, we maintain RandLA-Net++ and RandLA-Net3+ with the same loss function and other parameter settings as those in RandLA-Net, and we use the weighted cross-entropy loss function on all models. As shown in Table 1, the OA and mIoU of the RandLA-Net++ network are better than those of the RandLA-Net, with improvements of 2.1% and 4.4%, respectively; the OA and mIoU of the RandLA-Net3+ network are improved by 2.1% and 2.7%, respectively, compared with RandLA-Net. As far as specific objects are concerned, RandLA-Net++ and RandLA-Net3+ achieve an improvement from 0 to 10.7% and 6.9% in IoU index of rail class relative to the benchmark network RandLA-Net, respectively, which indicates that the shallow and deep semantic information and multi-scale features can improve the segmentation results of the datasets with unbalanced samples (features with less data) and complex scenes. At the same time, the segmentation accuracy of the overall ground objects is also improved to different degrees, which can effectively improve the semantic segmentation accuracy of the point clouds.

Table 1. Segmentation results evaluated for IoU in each category, mIoU, and OA of SensatUrban dataset. Note: the bold number represents the highest score in each column. The RandLA-Net network results is from this paper [20].

	OA	mIoU	Ground	Veg.	Building	Wall	Bridge	Parking	Rail	Traffic	Street	Car	Footpath	Bike	Water
RandLANet [20]	89.8	52.7	80.1	98.1	91.6	48.9	40.8	51.6	0.0	**56.7**	33.2	**80.1**	32.6	0.0	**71.3**
Ours (++)	**91.9**	**57.1**	**84.1**	**98.2**	94.4	58.6	**59.8**	53.4	10.7	54.6	**42.6**	78.2	38.2	0.0	69.7
Ours (3+)	**91.9**	55.4	83.7	**98.2**	**94.8**	50.9	53.9	**60.2**	6.9	56.1	39.1	76.8	**38.9**	0.0	61.4

4.3.2. Toronto-3D Dataset

The Toronto-3D dataset was acquired using the mobile laser system (MLS), which covers a point cloud of approximately 1 km and consists of approximately 78.3 million points and 8 labeled object classes, including roads, road markings, nature, buildings, utility lines, poles, cars, and fences. The dataset includes four regions: L001, L002, L003, and L004, of which the L002 region is used for testing, and the remaining three parts are used for training. The number of points in the training set is the combination of the L1, L3, and L4 regions, which contains approximately 68.0 million points, and the testing set is made up of the L2 region, which consists of approximately 10.3 million points.

Moreover, for a fairer comparison with the RandLA-Net network, we use the RandLA-Net++ network and the RandLA-Net3+ network with the same parameter settings and loss function as the RandLA-Net network. In the experiment, the mIoU and OA indicators are used to evaluate the segmentation results, as shown in Table 2. Compared with the benchmark network RandLA-Net, in this paper, RandLA-Net++ improves the mIoU indicator by 3.2%. The RandLA-Net3+ improves mIoU by 2.2% over the RandLA-Net network. From the perspective of specific categories, we achieve optimal or sub-optimal scores in most categories; especially for fences with less data, the IoU of fences is more than 10% higher than that of RandLA-Net, which again shows that our method is suitable for complex scenes, and that it has a better segmentation effect regarding small ground objects with unbalanced samples.

Table 2. Segmentation results evaluated on IoU of each category, mIoU, and OA of the Toronto3D dataset. Note: the bold number represents the highest score in each column. The RandLA-Net network results is from this paper [20].

	OA	mIoU	Road	Rd Mrk.	Natural	Building	Util. Line	Pole	Car	Fence
RandLA-Net [20]	93.0	77.7	94.6	42.6	**96.9**	93.0	86.5	78.1	**92.9**	37.1
Ours (++)	96.9	**80.9**	96.4	63.7	96.2	**94.8**	**86.8**	77.7	87.6	**43.6**
Ours (3+)	**97.0**	79.9	**96.8**	**70.0**	96.1	92.3	86.3	**80.4**	91.5	29.4

4.3.3. NJSeg-3D Dataset

The NJSeg-3D dataset is a point cloud dataset of urban scenes obtained by LiDAR, and it mainly annotates three main urban objects: buildings, trees, and roads. The NJSeg-3D dataset contains approximately 175.0 million points in the training dataset, 58.3 million points in the validation dataset, and 123.8 million points in the test dataset.

For the NJSeg-3D dataset, we keep the RandLA-Net++ network and the RandLA-Net3+ network with the same parameter settings and loss function as the RandLA-Net network, using the weighted cross-entropy loss function. It can be seen from Table 3 that under different feature combinations, the RandLA-Net++ network, based on shallow feature and deep feature fusion, achieves the highest overall accuracy and mIoU, along with its OA and mIoU, are 0.6% and 3.4% higher than that of the RandLA-Net, respectively; Moreover, the RandLA-Net3+ network OA and mIoU, based on multi-scale feature fusion, are also 0.5% and 3.2% higher, respectively, than those of the RandLA-Net. Regarding different objects, the prediction accuracy for buildings and trees is much better than the prediction results for roads. The main reason is that the number of road samples in the city is far less than the number of samples of buildings and trees. As with the test phase, it is more inclined to predict the outcome of buildings and trees. For buildings, although the spectral information is similar to that of roads, buildings, as ground objects with significant three-dimensional spatial features, have a better overall segmentation effect. For buildings, after multi-feature fusion, the false detections and missed detections are improved, to a certain extent. In summary, the feature fusion of different depths or different scales is beneficial to the semantic segmentation of 3D point clouds, especially in the segmentation, false detection, and missed detection of complex and confusing objects.

Table 3. Segmentation results evaluated on IoU of each category, mIoU, and OA of the NJSeg-3D dataset. Note: the bold numbers represent the highest score in each column.

	Type	OA%	Recall %	Accuracy%	F1	IoU%	mIoU%
RandLA-Net [20]	building	94.5	93.5	86.0	91.5	80.1	82.1
	road		82.4	75.4	78.7	65.0	
	tree		91.5	97.5	95.5	91.4	
	others		96.3	**95.8**	96.1	92.4	
RandLA-Net++	building	**95.1**	**96.5**	**91.9**	**94.7**	**89.9**	**85.5**
	road		**84.3**	**79.5**	**81.8**	**69.3**	
	tree		91.9	96.5	95.2	90.1	
	others		97.6	94.9	96.2	92.7	
RandLA-Net3+	building	95.0	94.9	90.2	93.5	87.8	85.3
	road		83.2	75.9	79.4	65.8	
	tree		**96.9**	**98.8**	**96.8**	**93.8**	
	others		**97.8**	95.7	**96.7**	**93.7**	

We list some prediction samples of local areas in Figure 8, and Figure 9 is a partially enlarged view, where each column represents a different point cloud example, where Point Cloud is the original point cloud, Label is the real label, and RandLA-Net, RandLA-Net++, and RandLA-Net3+ represent semantic segmentation networks with different feature combinations, respectively. It can be seen from the figure that the method based on multi-feature fusion shows a significant improvement regarding correct and false detection, as well as missed detection, and the boundary of the ground objects is finer and more accurate. These details reveal the superior performance of our method.

4.3.4. Efficiency of RandLA-Net++ and RandLA-Net3+

In this section, we systematically evaluate the overall efficiency of our RandLA-Net++ and RandLA-Net3+ regarding real-world large-scale point clouds for semantic segmentation. Particularly, we evaluate RandLA-Net++ and RandLA-Net3+ using the SensatUrban [11] dataset, obtaining the total time consumption of our network on the test set, i.e., sequences 2 and 8 of the Birmingham urban area, and sequences 15, 16, 22, and 27 of the Cambridge urban area. For a fair comparison, we feed the same number of points (i.e., 65,536) from each scan into each neural network. In addition, we also evaluate the memory consumption and model training time of RandLA-Net++ and RandLA-Net3+. Note that all experiments are conducted on the same machine with an Intel i9-10980XEX @3.0GHz CPU and an NVIDIA Quadro RTX 6000 GPU.

Table 4 quantitatively shows the memory consumption, training time, and inference time of different approaches. Compared with the RandLA-Net network, our improved networks, RandLA-Net++ and RandLA-Net3+, show a significant increase in the number of parameters, training time, and inference time, indicating that our improved model has a slight increased the complexity over the other models. In future work, we will consider how to further reduce the complexity of the model through methods such as model pruning.

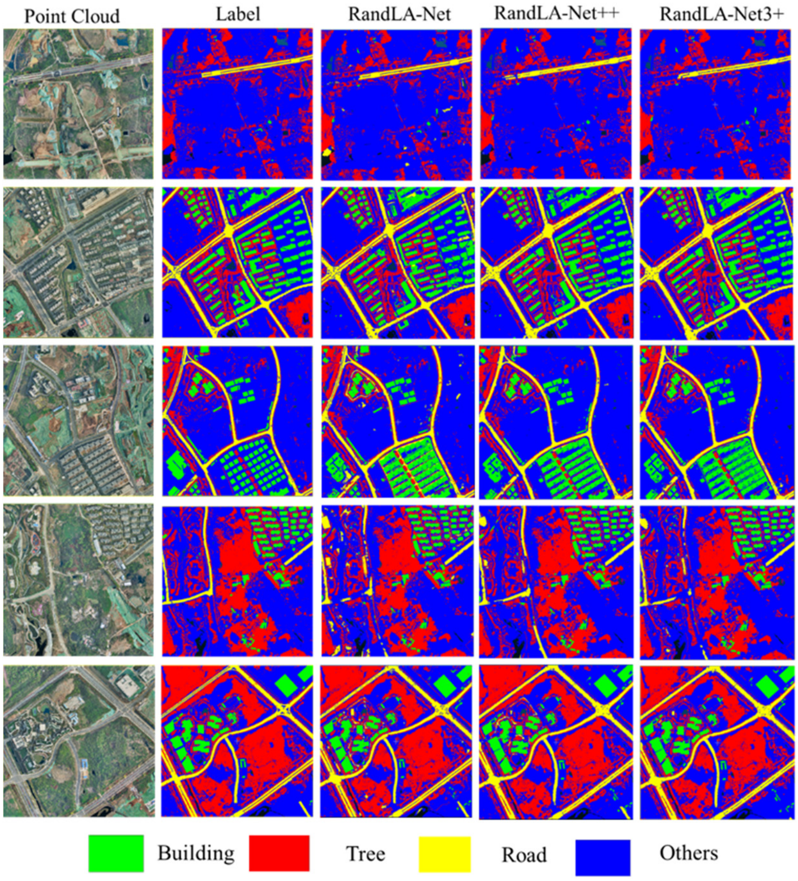

Figure 8. Point cloud semantic segmentation results.

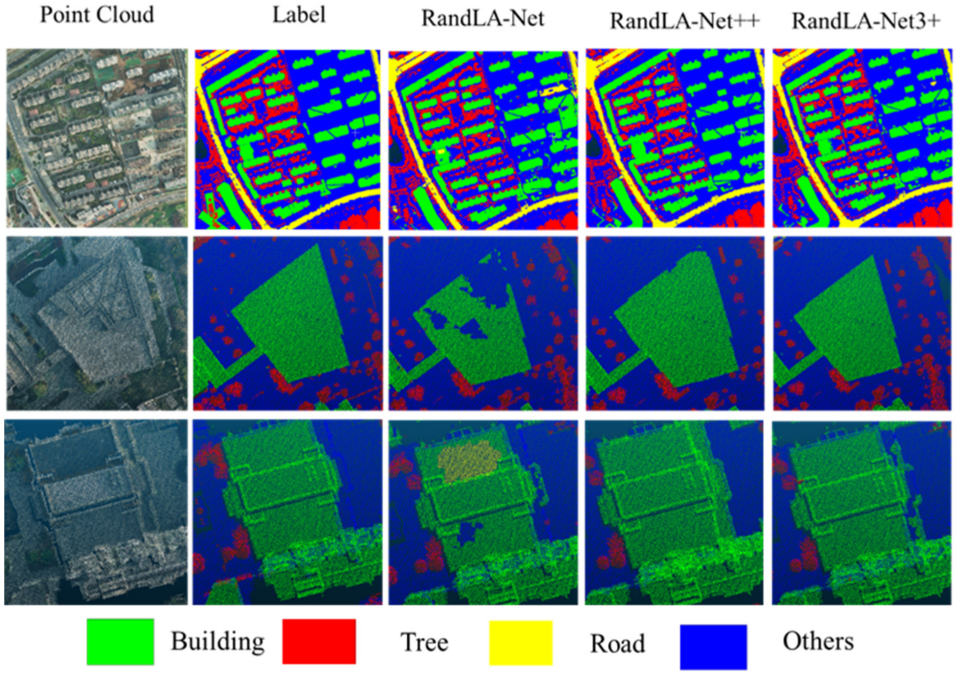

Figure 9. Partially enlarged view of the point cloud semantic segmentation results.

Table 4. The network parameters, training time, and inference time of different approaches for semantic segmentation using the SensatUrban [11] dataset.

	Parameters (Millions)	**Training Time (Hours)**	**Inference Time (Seconds)**
RandLA-Net	14.9	16 ± 1	362 ± 5
RandLA-Net++	17.6	20 ± 1	371 ± 5
RandLA-Net3+	18.6	18 ± 1	375 ± 5

4.3.5. The Impact of Imbalanced Class Distribution

To alleviate the imbalanced class distribution problem, a typical solution is to use more sophisticated loss functions. We evaluate the effectiveness of three off-the-shelf loss functions based on the NJSeg-3D dataset, using RandLA-Net++ as a baseline network. The loss functions are: weighted cross-entropy loss, Lovász-Softmax loss, and the combined loss of weighted cross-entropy and Lovász-Softmax loss. In the NJSeg-3D dataset, the sample sizes of roads and buildings are small, accounting for 10% and 20% of the total samples, respectively. As shown in Table 5, after using the Lovász-Softmax loss function, the IoU scores for roads and buildings were improved by 1.6% and 3.3% over the baseline model, respectively; after using a linear combination of weighted cross-entropy loss and Lovász-Softmax loss, the IoU scores for roads and buildings were improved by 2.1% and 3.2% over the baseline model, respectively, and its mIoU reached 86.8%, which indicates that the loss function can improve the imbalanced class distribution problem.

Table 5. Segmentation results evaluated on IoU of each category, mIoU, and OA of the NJSeg-3D dataset. Whether to add dilated convolution and different loss functions, respectively. Note: the bold numbers represent the highest score in each column.

	OA	**mIoU**	**Building**	**Tree**	**Road**	**Others**
RandLA-Net (L_{wce})	94.5	82.1	80.1	**91.4**	65.0	92.4
RandLA-Net++ (L_{wce})	95.1	85.5	89.9	90.1	69.3	92.7
RandLA-Net++ (L_{ls})	95.8	86.2	**93.2**	86.7	70.9	93.8
RandLA-Net++ (L_{wce}+ L_{ls})	**96.0**	**86.8**	93.1	88.5	**71.4**	**94.0**

5. Discussion

For the point cloud semantic segmentation of typical features (such as buildings, trees, roads, etc.) in urban scenes, the RandLA-Net++ and RandLA-Net3+ networks are proposed by fusing features of different depths and scales based on RandLA-Net. The main contributions include the following three points. First, the optimal depth selection of the model is based on the fusion of the deep and shallow features of the point clouds. Second, multi-scale feature fusion is used to capture full-scale fine-grained details and coarse-grained semantic information. Finally, we propose to incorporate dilated convolution to increase the receptive field and compare the improvement effect of different loss functions on sample class imbalance. In short, the proposed method shows great improvement regarding the missed detection and false detection of buildings and other objects, the semantic segmentation of complex scenes, and the category imbalance, especially for categories with fewer data. Although based on RandLA-Net and its improved algorithm, the typical features of urban scenes have obtained better segmentation results, but there is still a problem regarding the boundary accuracy of buildings, roads and other features. In subsequent research, we can consider adding features such as normal vectors in order to enhance the detection ability of regular features, including buildings. In addition, the generalization of the 3D point cloud semantic segmentation model to different urban datasets is also of great research significance.

Author Contributions: Conceptualization, J.C. and Y.Z.; methodology, J.C. and Y.Z.; writing—original draft preparation, J.C. and Y.Z.; writing—review and editing, J.C., C.M., Y.L. and Y.Z.; supervision, Y.Z.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Open Fund of the Key Laboratory of Degraded and Unused Land Consolidation Engineering, Ministry of Natural Resources (No. SXDJ2019-4); the Scientific Re-search Project of Nanjing Institute of Surveying, Mapping & Geographical Investigation, CO. Ltd. (No. H7P210062); and the Natural Science Foundation of China (No. 51874306).

Data Availability Statement: The data presented in this study are available in SensatUrban dataset [11] (http://point-cloud-analysis.cs.ox.ac.uk/ (accessed on 1 September 2022)) and Toronto-3D dataset [12] (https://github.com/WeikaiTan/Toronto-3D (accessed on 1 September 2022)). In addition, the NJSeg-3D dataset are available on request from the corresponding author. The NJSeg-3D dataset are not publicly available due to privacy restrictions.

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

References

1. Kasznar, A.P.P.; Hammad, A.W.; Najjar, M.; Linhares Qualharini, E.; Figueiredo, K.; Soares, C.A.P.; Haddad, A.N. Multiple dimensions of smart cities' infrastructure: A review. *Buildings* **2021**, *11*, 73. [CrossRef]
2. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
3. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
4. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
5. Yang, B.; Liang, F.; Huang, R. Progress, challenges and perspectives of 3D LiDAR point cloud processing. *Acta Geod. Cartogr. Sin.* **2017**, *46*, 1509–1516.
6. Liao, X. Scientific and technological progress and development prospect of the earth observation in China in the past 20 years. *Natl. Remote Sens. Bull.* **2021**, *25*, 267–275.
7. Yi, L.; Su, H.; Guo, X.; Guibas, L.J. SyncSpecCNN: Synchronized spectral CNN for 3D shape segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6584–6592.
8. Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; Xiao, J. 3DShapeNets: A deep representation for volumetric shapes. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1912–1920.
9. Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Nießner, M. ScanNet: Richlyannotated 3D reconstructions of indoor scenes. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2432–2443.
10. Hackel, T.; Savinov, N.; Ladicky, L.; Wegner, J.D.; Schindler, K.; Pollefeys, M. SEMANTIC3D.NET: A new large-scale point cloud classification benchmark. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *I V-1/W1*, 91–98. [CrossRef]
11. Hu, Q.; Yang, B.; Khalid, S.; Xiao, W.; Trigoni, N.; Markham, A. Towards semantic segmentation of urban-scale 3D point clouds: A dataset, benchmarks and challenges. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Virtual, 19–25 June 2021; pp. 4977–4987.
12. Tan, W.; Qin, N.; Ma, L.; Li, Y.; Du, J.; Cai, G.; Yang, K.; Li, J. Toronto-3D: A large-scale mobile LiDAR dataset for semantic segmentation of urban roadways. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 202–203.
13. Yang, B.; Dong, Z. Progress and perspective of point cloud intelligence. *Acta Geod. Cartogr. Sin.* **2019**, *48*, 1575–1585.
14. Yu, B.; Dong, C.; Liu, Y. Deep learning based point cloud segmentation: A survey. *Comput. Eng. Appl.* **2020**, *56*, 38–45.
15. 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]
16. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-view convolutional neural networks for 3d shape recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 945–953.
17. Maturana, D.; Scherer, S. Voxnet: A 3d convolutional neural network for real-time object recognition. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–3 October 2015.
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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
19. 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 the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.

20. 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), Seattle, WA, USA, 13–19 June 2020; pp. 11108–11117.
21. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. Pointcnn: Convolution on x-transformed points. *Adv. Neural. Inf. Process. Syst.* **2018**, *31*, 820–830.
22. 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 (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 6411–6420.
23. Wang, C.; Samari, B.; Siddiqi, K. Local spectral graph convolution for point set feature learning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 52–66.
24. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.D.; Solomon, J.M. Dynamic Graph CNN for Learning on Point Clouds. *Acm Trans. Graph.* **2019**, *38*, 1–12. [CrossRef]
25. Himmelsbach, M.; Hundelshausen, F.V.; Wuensche, H.J. Fast segmentation of 3D point clouds for ground vehicles. In Proceedings of the 2010 IEEE Intelligent Vehicles Symposium, La Jolla, CA, USA, 21–24 June 2010; pp. 560–565.
26. Liu, F.; Li, S.; Zhang, L.; Zhou, C.; Ye, R.; Wang, Y.; Lu, J. 3DCNN-DQN-RNN: A deep reinforcement learning framework for semantic parsing of large-scale 3D point clouds. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5678–5687.
27. Zhao, C.; Zhou, W.; Lu, L.; Zhao, Q. Pooling scores of neighboring points for improved 3D point cloud segmentation. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1475–1479.
28. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 1856–1867. [CrossRef]
29. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–8 May 2020; pp. 1055–1059.
30. Pang, Y.; Li, Y.; Shen, J.; Shao, L. Towards bridging semantic gap to improve semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4230–4239.
31. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Jorge Cardoso, M. Generalised dice overlap as a deep learning loss function for highly unbalanced seg-mentations. In *Deep Learning in Medical Image Analysis and Multi-Modal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2017; pp. 240–248.
32. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 2980–2988.
33. Cortinhal, T.; Tzelepis, G.; Aksoy, E.E. Salsanext: Fast semantic segmentation of lidar point clouds for autonomous driving. *arXiv* **2020**, arXiv:2003.03653.