ISSN: 2472-3703 DOI: 10.6911/WSRJ.202305_9(5).0007

DRandLA-Net: A Point Cloud Classification Model for Large-Scale Photogrammetry in Complex Scenes

Caoyuan Song, Haiyang Yu*, Zhihua Hua, Saifei Xie and Peng Jing

School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo, 454003, China

Abstract

In this paper, we propose an enhanced model for large-scale point cloud semantic segmentation based on the RandLA-Net neural network architecture. Our model aims to capture both fi-ne-grained details and coarse-grained semantic information by deepening the depth of the dilated residual structure, redesigning the connections between the encoder and decoder, and improving the internal connections between the decoders. This leads to the creation of a three-dimensional point cloud semantic segmentation neural network model, DRandLA-Net, specifically designed for large-scale photogrammetric measurement. Our model achieves an overall accuracy of 90.97% and 87.42% on public and self-built datasets, respectively, with an average intersection-over-union (IoU) of 53.84% and 42.54%, respectively. Compared to the segmentation results of the RandLA-Net model, our model improves the overall accuracy by 1.19% and 2.28%, and the average IoU by 1.15% and 2.94%, respectively. Additionally, our model demonstrates excellent generalization ability.

Keywords

Photogrammetric point cloud; Semantic segmentation; Deep learning; DRandLA-Net.

1. INTRODUCTION

In recent years, tilt photogrammetry technology has made significant strides and has become popular due to its high efficiency and cost-effectiveness [1]. The acquired point cloud data contains valuable 3D position information of the target as well as rich spectral information. For achieving real scene simulation and analysis, intelligent se-mantic segmentation of large scene photogrammetry point clouds is crucial for 3D re-construction and 3D digital mapping. It also plays a vital role in 3D scene under-standing. Efficient and accurate point cloud segmentation methods are in high de-mand due to the rapid development of tilt photogrammetry technology. The proposed DRandLA-Net model provides an effective solution for semantic segmentation of large-scale photogrammetric point cloud data, contributing to the development of 3D digital mapping, 3D scene understanding, and other related fields with a focus on sustainability.

Currently, 3D point cloud semantic segmentation methods can be classified into two categories: traditional point cloud semantic segmentation methods and deep learning-based point cloud semantic segmentation methods [2-6]. Traditional shallow machine learning-based point cloud segmentation methods generally have low accu-racy. On the other hand, deep learning-based 3D point cloud semantic segmentation methods can be divided into two subcategories, namely indirect segmentation methods and direct segmentation methods. Indirect segmentation methods include voxel-based methods and multi-view based methods. However, voxel-based methods can consume substantial memory resources, while the segmentation accuracy of multi-view based methods may be impacted by information loss due

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202305_9(5).0007

to projection dimensionality reduction. Direct segmentation methods refer to processing point clouds directly through designing neural network models, such as the MVCNN [7] and SnapNet [8] algorithms. While these methods do not require conversion of the topological relationship of point clouds, they have relatively high computational complexity. In [9], PointNet, a groundbreaking end-to-end neural network for 3D point clouds was proposed. The network uses a shared Multilayer Perceptron (MLP) to learn single point features in the point cloud and solves the problem of rotation invariance using a transformation network (T-net). However, this method did not consider the local features between points in the point cloud, limiting its segmentation ability for complex scenes. To ad-dress this, Qi et al. [10] improved the PointNet network by using hierarchical point cloud feature learning to extract local features between points. Additionally, the literature aggregate local point cloud features by accumulating information from previous layer networks [11-13], while [14] aggregate local features of point clouds based on attention mechanisms. These algorithms perform well in point cloud semantic seg-mentation tasks in small scenes, but their performance is limited in large scenes due to high memory and hardware requirements. To address the limitations of existing methods, Huqingyong et al. [15] proposed RandLA-Net, a large-scale and efficient se-mantic segmentation network for point clouds. The algorithm designs a local feature aggregation module to learn features between points and combines it with a random sampling module to reduce processing time for point clouds, achieving good results in large-scale point cloud segmentation tasks. Chenyi et al. [16] proposed an improved neural network structure of RandLA-Net to construct a deeper 3D point cloud seman-tic segmentation neural network model for large-scale unstructured agricultural scenes. The experimental results showed good performance. Xuhan et al. [17] proposed an efficient point cloud spatial downsampling sampling strategy combined with spa-tial aggregation to effectively encode local features. They constructed an end-to-end network for semantic segmentation of urban scenes, addressing the imbalance problem of different classes by loss function. Jingdu et al. [18] proposed a large-scale point cloud semantic segmentation network (ResDLPS-Net) without chunking operation. They designed a novel feature extraction module to efficiently extract neighborhood, geometric, and semantic features. Through an attention mechanism, the learned fea-tures were aggregated to form local feature descriptors. In literature [19], Rand-LA-Net++ and RandLA-Net3+ semantic segmentation networks were proposed based on the RandLA-Net model. They improved the semantic segmentation accuracy of point clouds by employing deep fusion of shallow and deep features of point clouds using jump connections to fully capture the detailed features between point clouds. Although these models have shown good performance in segmenting large-scale point clouds, they still need to improve their ability to segment smaller ground classes in the scene and to extract detailed features. This paper proposes the DRandLA-Net model, which combines a deep local feature aggregation module and a multi-scale fusion module. It comprehensively captures fine-grained details and coarse-grained semantic information and improves the feature correlation between points. It has better seg-mentation performance for complex scenes and small land objects with imbalanced samples. The proposed model provides an effective solution for semantic segmentation of large-scale photogrammetric point cloud data, contributing to sustainable development in related fields.

In summary, this paper proposes the DRandLA-Net model, which is based on the RandLA-Net algorithm. The model includes a deep local aggregation module to im-prove the network's capability to extract detailed features. The model also uses jump connections to fuse multi-scale features and build an end-to-end neural network for semantic segmentation of 3D point clouds for large scene photogrammetry. The pro-posed model achieves good semantic segmentation results with improved capability to extract detailed features. The segmentation effect of the proposed method on different feature point clouds is analyzed through validation experiments. The DRandLA-Net model provides an effective and feasible solution for semantic segmentation

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202305_9(5).0007

of large-scale photogrammetric point cloud data, contributing to sustainable development in related fields.

2. MATERIALS AND METHODS

2.1. Study Area

The experimental data used in this study were obtained from the point cloud data generated by the oblique photogrammetric images of a test area collected by the Huace P330 Pro aerial survey drone. The data collection process included five data collection plans, with a 20% overlap at each site and an average ground sampling distance of 5.44 cm. The resulting sampling pixel was 7952×5304. Analytical aerial triangulation was performed using modeling software to generate the oblique photogrammetric point cloud data. The data were manually annotated using point cloud processing software, as shown in Figure 1. The dataset contains nearly 400 million points with detailed semantic annotations, and each point is labeled as one of nine semantic categories, including ground, vegetation, building, wall, road, road furniture, car, sidewalk, and water. Corresponding values of 0, 1, 2, 3, 4, 5, 6, 7, and 8 were set for these nine la-bels during training. The data were divided into 18 blocks, with a ratio of 14:2:2 for the training set, validation set, and test set, respectively.

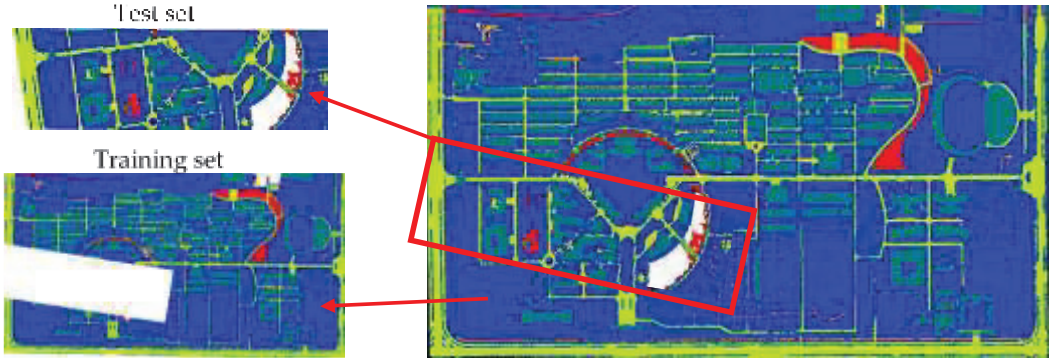


Figure 1. HPU data set

2.2. RandLA-Net model main module components

2.2.1. Random downsampling

Random Sampling (RS) is a commonly used point cloud downsampling method, which uniformly selects K points from the input N points, each with the same probability of being selected. The computational effort of downsampling is independent of the total number of input point clouds, and is only related to the number of points K after downsampling, constant complexity. The RS algorithm can be expressed as:

$$RS(X) = S \mid S = X[i] \tag{1}$$

Where $i \in [1, N]$, where X is the input data, N is the number of points in the input data, and S is a set of sampling points.

Random downsampling can effectively reduce the processing time of point clouds and save the cost of hardware facilities and time cost, but it may lose some important geometric structure features, which affects the accuracy of point cloud classification.

2.2.2. Local feature aggregation module

The local feature aggregation module comprises three main components: Local Spatial Encoding (LocSE), Attentive Pooling (AP), and Dilated Residual Block (DRB), as depicted in Figure 2.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202305_9(5).0007

The LocSE module is designed to encode the three-dimensional coordinate information of the input point cloud. Unlike the conventional method of directly inputting the three-dimensional coordinates of each point as a regular channel feature in the network, the LocSE module explicitly encodes the spatial geometric information of the three-dimensional point cloud. This allows the network to better understand the spatial geometry structure by leveraging the relative positions and distance information of each point. The module executes the following steps:

Firstly, the K-nearest neighbor (KNN) search algorithm is used to find the nearest neighboring points in the Euclidean space for each point. For the nearest neighboring points $\{P_i^1 \ldots P_i^2 \ldots P_i^k\}$ of point P_i, their relative positions are explicitly encoded by concatenating the three-dimensional coordinates of the center point P_i, the neighboring point P_i^k, the relative coordinates $(P_i - P_i^k)$, and the Euclidean distance $\| P_i - P_i^k \|$ together. As follows:

$$r_i^k = MLP(p^i \oplus p_i^k \oplus (p^i - p_i^k) \oplus \| p^i - p_i^k \|) \tag{2}$$

In the formula, P_i represents the three-dimensional coordinates of the center point; P_i^k represents the three-dimensional coordinates of the neighboring points; r_i^k represents the encoded relative positions of the center point and neighboring points. Finally, the point features f_i^k corresponding to the neighboring points P_i^k are concatenated with the encoded relative positions r_i^k to obtain the new point features f_i^k.

The attention pooling module is designed to automatically learn and aggregate useful information from the feature sets of neighboring points through attention mechanisms. For a set of neighboring feature points $\hat{F}_i^k = \{\hat{f}_i^1 \ldots \hat{f}_i^2 \ldots \hat{f}_i^k\}$ in a point cloud, a shared function $g(.)$ is designed to learn a separate attention score for each point. As follows:

$$S_i^k = g(\hat{f}_i^k, W) \tag{3}$$

In the formula, W represents the learnable parameters of the shared MLP; $\hat{f}_i^k$ represents the feature of neighboring points; and S_i^k represents the learned attention score.

The learned attention scores are regarded as a mask that can automatically select important features, and the resulting feature is a weighted sum of the feature points in the neighborhood, with the weights determined by the learned attention scores, As follows:

$$\tilde{f}_i = \sum_{k=1}^{k} (\hat{f}_i^k . S_i^k) \tag{4}$$

In the formula, $\tilde{f}_i$ represents the weighted sum of the feature points in the neighborhood, and k represents the number of neighboring points.

Expansion Residual Block is formed by combining two sets of LocSE, Attention Pooling, and skip connections. This deep Expansion Residual Block is also known as the Local Feature Aggregation Module (LFA), which effectively increases the receptive field and promotes feature propagation between neighboring points by aggregating features at each step. As a result, this approach provides a more cost-effective way to perform feature aggregation.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202305_9(5).0007

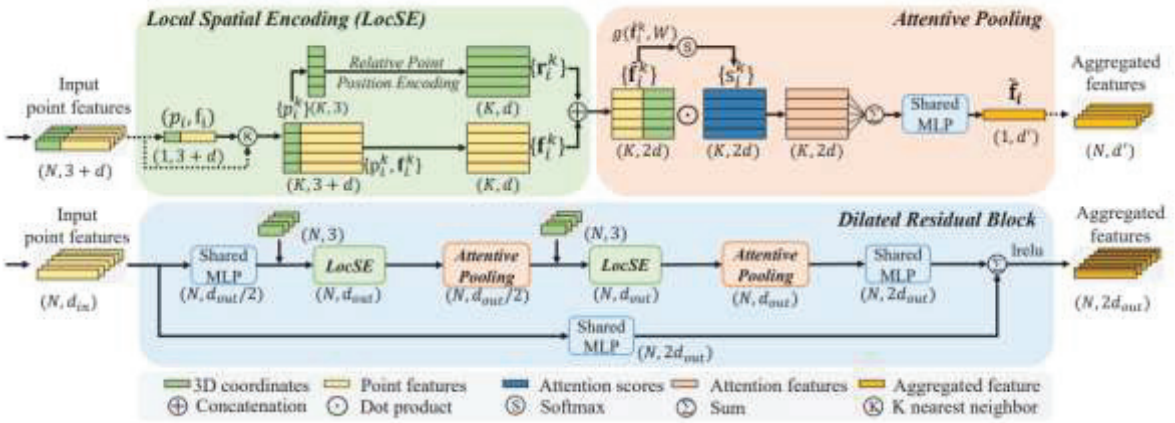


Figure 2. Local feature aggregation module

2.3. DRandLA-Net model

2.3.1. Deep local feature aggregation module

In order to enhance the network's ability to extract local features from point clouds, we designed a Deep Local Feature Aggregation (DLFA) module, as illustrated in Figure 3. This module is based on the local feature aggregation module of RandLA-Net, and deepens the Expansion Residual structure by incorporating three sets of LocSE and Attention Pooling. We also made adjustments to the multi-layer perceptron (MLP) and skip connections. The MLP is used for fusing the outputs from different layers, which increases the range of feature fusion and improves the model's ability to capture detailed information. The resulting module is called the Deep Dilated Residual Block (DDRB), which efficiently increases the receptive field and promotes feature propagation between neighboring points by aggregating features at each step. The DDRB achieved impressive results in the segmentation of large-scale point clouds. Furthermore, to improve the segmentation accuracy of small target classes, we assigned them a larger weight to enhance the network's ability to extract features of these objects.

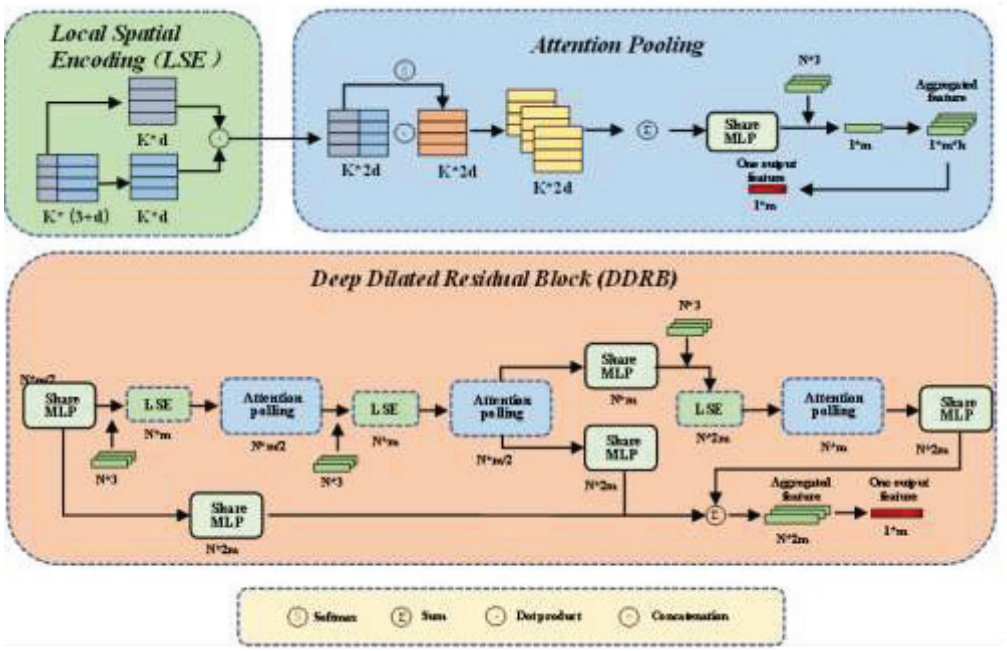


Figure 3. Deep local feature aggregation module

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202305_9(5).0007

2.3.2. Jump connection

Jump connections were first applied to classification and segmentation networks of images. They can effectively combine semantic information in high-level features and spatial information in low-level features to optimize the segmentation effect. Long et al. [20] first introduced jump connections to fully convolutional networks by adding the feature mapping of upsampling to the feature mapping of the encoder. Unlike the common jump connections in U-Net [21], Zhou et al. [22] proposed a new segmentation structure based on nested and dense jump connections, U-Net++. This structure uses dense jump connections to achieve semantically different feature mapping fusion for more efficient capture of foreground objects and the fine-grained details of the foreground objects. The jump connections help in parameter updating and can solve the problem of gradient disappearance in deeper network layers. This reduces the time loss of the net-work learning process and speeds up the convergence of the network.

To fuse multi-scale features, we utilize full-size skip connections between the encoder and decoder and connect multi-scale features within the decoder. In the encoder, we downsample the features of each layer through max-pooling to obtain the corresponding scale features in the decoder, which are then fused. In the decoder, the out-put features of each layer are interpolated to the corresponding scale using near-est-neighbor interpolation. Finally, these features are concatenated to fuse the multiscale features. This approach significantly improves the accuracy of point cloud se-mantic segmentation.

2.3.3. Neural Network Architecture Design

During the encoding and decoding process of the network, we adjusted various parameters, such as the ratio of random downsampling and the feature dimension size of each layer, to achieve optimal segmentation accuracy on the dataset. The entire network structure comprises an input layer, encoding layers, decoding layers, and an output layer, all of which significantly improve the accuracy of point cloud semantic segmentation. Figure 4 illustrates the network structure, the network structure consists of four main parts: the input layer, encoding layers, decoding layers, and output layer.

The input layer receives the 3D point cloud, where each point in the dataset contains seven attribute information: X, Y, Z, R, G, B, and L, representing the position in-formation X, Y, Z, color R, G, B, and label category L.

The encoding layer includes a deep local feature aggregation module and a random sampling module (RS). In the encoding part of the network, we use a random sampling layer and four encoding layers to gradually reduce the size of the input point cloud data. The downsampling ratio of the five layers is set to 6:6:6:6:4, with an in-crease in the feature dimension of the sampled points at each layer. The data volume of the point cloud is reduced to 1/6 of the previous layer after each of the first four layers, and the last layer is reduced to 1/4 of the previous layer. The feature dimensions of the sampled points after downsampling are (8, 32, 128, 256, 512), respectively.

The decoding layer consists of an upsampling layer (US) and a shared multilayer perceptron (MLP). We use four decoding layers to gradually reduce the feature dimensions of each point and restore the number of point clouds to the size of the input. After four decoding layers, the feature dimensions of the sampled points output are (512, 256, 128, 32, 8), respectively.

The last three layers of the neural network use fully connected layers (FC) to map the learned features to the sample label space. In this network, the fully connected layers are used to connect the hidden layers and the output layers for subsequent processing. To prevent overfitting, a Dropout layer is added to the end of the hidden layer, with a keep rate of 0.7, to prevent overfitting during the training process and enhance the generalization ability of the neural network.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202305_9(5).0007

In addition, skip connections are used to concatenate the features of the encoding side to obtain multi-scale fusion point cloud features. Full-size skip connections are used between the encoder and decoder, and multi-scale features are connected within the decoder. Specifically, in the encoder, we downsample the features of each layer through max-pooling to obtain the corresponding scale features in the decoder, and then fuse them. In the decoder, the output features of each layer are interpolated to the corresponding scale using nearest-neighbor interpolation. Finally, these features are concatenated to fuse the multi-scale features. This approach improves the accuracy of point cloud semantic segmentation by capturing features at different scales.

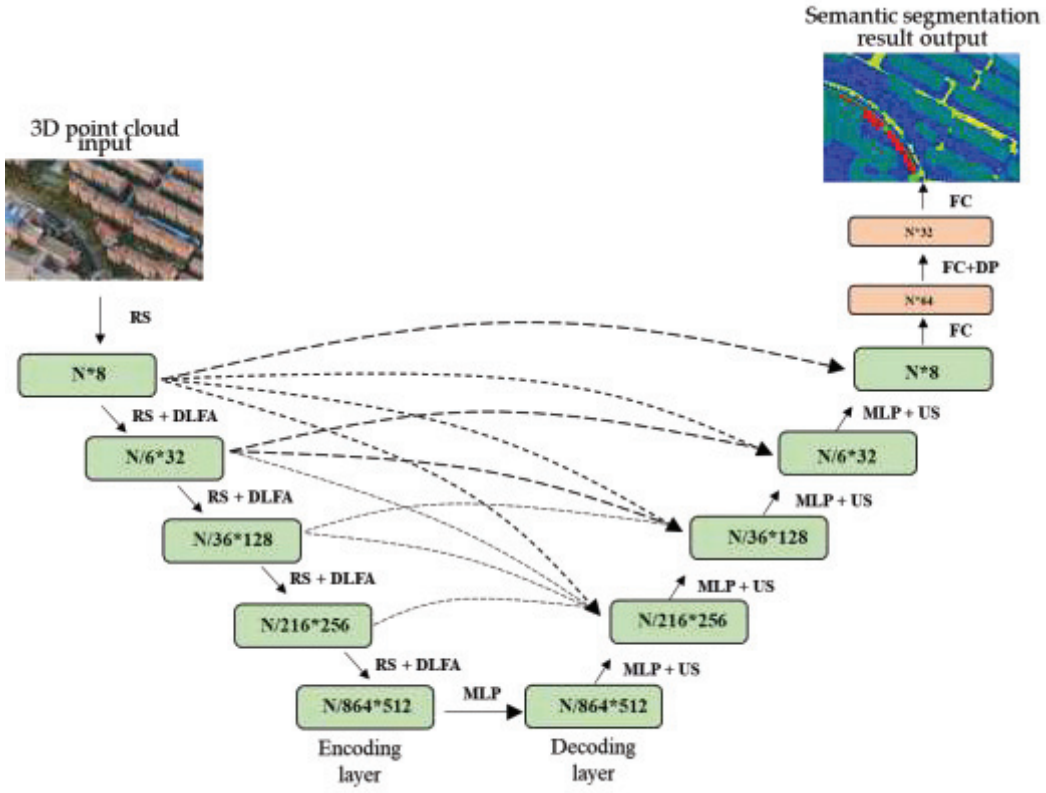


Figure 4. Neural network architecture (N*D denotes the number of points and feature dimension, respec-tively; FC: fully connected layer; DLFA: deep local feature aggregation; RS: random sampling; MLP: shared multilayer perceptron; US: upsampling; DP: dropout)

2.4. Experimental environment and parameter configuratio

The experiment was conducted on a system running Ubuntu 16.04 with a single RTX 2080Ti graphics card, 64 GB of RAM, and an Intel Core i9-9900K processor. The deep learning framework used was TensorFlow 1.11, with programming language Python 3.6. GPU acceleration was achieved using CUDA 9.0 and cudnn 7.4.5. The training process took approximately 17 hours.

To prepare the training data, the point cloud data were first downsampled using a grid size of 0.2 m, and the color was normalized (RGB/255). A Kdtree was construct-ed using the coordinates of the downsampled point cloud, and the index of the nearest downsampled point for each original point was queried using the Kdtree with a value of 16 for the number of nearest points. The resulting list of indices, which is equivalent in size to the number of original points, was saved.

During training, the network was trained with a batch size of 4 and a learning rate of 0.01, and the number of points in each batch was set to 65536. The decay rate was set to 0.95, and the maximum Epoch for training was set to 100.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202305_9(5).0007

2.5. Precision evaluation index

To evaluate the performance of the semantic segmentation model, two metrics were used in the paper: Mean Intersection over Union (MIoU) and Overall Accuracy (OA). The formula is as follows:

$$OA = \frac{TP + TN}{TP + FN + FP + TN} \tag{5}$$

$$MIoU = \frac{1}{k}\sum_{i=0}^{k} \frac{TP}{TP + FN + FP} \tag{6}$$

In these metrics, TP represents true positive (both the true label and the predicted label are 1), TN represents true negative (both the true label and the predicted label are 0), FP represents false positive (the true label is 0 while the predicted label is 1), FN represents false negative (the true label is 1 while the predicted label is 0), and k is the number of classes.

3. RESULTS

3.1. Results of classification experiments on SensatUrban dataset

In this section, the segmentation experiments were conducted using the SensatUrban [23] dataset of urban-scale point clouds shared by the University of Oxford. The training data was divided into training set, validation set, and test set in a ratio of 30:7:5. In the experiment, regions 1, 5, 7, and 10 were used as the validation set, while regions 2, 8, 15, 16, 22, and 27 were used as the test set to evaluate the DRandLA-Net segmentation network model and compare it with other segmentation network models. The accuracy values are shown in Table 1.

Table 1. Segmentation accuracy evaluation of different segmentation methods

Models	IoU(%)						
	Grounds	Vegetation	Buildings	Walls	Bridges	Parking	Railroad tracks
CloserLook3D	55.13	90.54	78.24	**54.06**	13.11	14.62	0
RandLA-Net	80.11	98.07	91.58	48.88	40.75	51.62	0
PointNet++	72.46	94.24	84.77	2.72	2.09	25.79	0
BAF-LAC	73.55	95.73	92.74	44.39	5.95	47.28	0
PointNet	67.96	89.52	80.05	0	0	3.95	0
Our	**80.16**	**98.59**	**93.07**	49.54	**41.05**	**52.21**	0
Models	Traffic ways	Street facilities	Cars	Sidewalk	Bicycles	Water	
CloserLook3D	37.83	24.47	74.24	14.91	0	10.94	
RandLA-Net	56.67	33.23	80.14	32.63	0	71.31	
PointNet++	31.54	11.42	38.84	7.12	0	56.93	
BAF-LAC	51.68	30.99	74.62	19.44	0	28.05	
PointNet	31.55	0	35.14	0	0	0	
Our	**57.03**	**35.45**	**81.15**	**39.48**	0	**72.26**	

The experimental results in Table 1 show that the proposed DRandLA-Net meth-od achieves better IoU than the RandLA-Net algorithm. The point clouds of ground, vegetation, buildings,

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202305_9(5).0007

and roads have wide coverage and clear structural features, and can all be predicted with relatively high segmentation accuracy. Compared to the four methods of CloserLook3D, PointNet++, BAF-LAC, and PointNet, the proposed method also achieves good segmentation accuracy for various classes. However, for small-sized point clouds of classes such as walls, cars, and sidewalks, it is difficult to fully learn their structural features using only the local feature aggregation module. The proposed DRandLA-Net method constructs a deep local feature aggregation mod-ule to expand the receptive field. Meanwhile, skip connections are used to connect be-tween the encoder and decoder, and internal connections are used between the decoders to achieve multi-scale feature fusion, which comprehensively captures fine-grained information and coarse-grained semantic information, enhances the correlation be-tween points, and effectively improves the semantic segmentation ability of 3D scenes. The experimental results show that the proposed method is suitable for complex scenes and has better segmentation performance for small objects with imbalanced samples.

Table 2. Segmentation results on the SensatUrban dataset

Models	MIoU(%)	OA(%)
CloserLook3D	36.01	80.44
RandLA-Net	52.69	89.78
PointNet++	32.92	84.30
BAF-LAC	43.41	86.57
PointNet	23.71	80.78
Our	**53.84**	**90.97**

From Table 2, it can be seen that the DRandLA-Net algorithm achieves an overall accuracy (OA) of 90.97% and a mean intersection-over-union (mIoU) of 53.84% in scene segmentation. Compared to the PointNet model that directly classifies point clouds, DRandLA-Net considers the feature relationships between points, resulting in an improvement of 10.19% in overall accuracy and 30.13% in mIoU. Compared to the PointNet++ segmentation model, DRandLA-Net algorithm has a higher segmentation efficiency while improving the segmentation accuracy, with an increase of 6.67% in overall accuracy and 20.92% in mIoU. Compared to large-scale point cloud classification networks such as BAF-LAC, CloserLook3D, and RandLA-Net, DRandLA-Net can comprehensively capture fine-grained details and coarse-grained semantic information, enhance the correlation between points, and improve segmentation accuracy, with an increase in overall accuracy of 4.4%, 10.53%, and 1.19%, and an increase in mIoU of 10.43%, 17.83%, and 1.15%, respectively.

Figure 5 shows the visualization results of the test set in the SensatUrban dataset. From the marked black ellipse in Figure 5, it can be seen that the DRandLA-Net algorithm has more accurate point cloud segmentation results for narrow sidewalks, cars, and small buildings next to the street. CloserLook3D and RandLA-Net algorithms have misclassification problems in segmenting sidewalks and roads, and they tend to mis-classify cars and small buildings next to each other as street facilities.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202305_9(5).0007

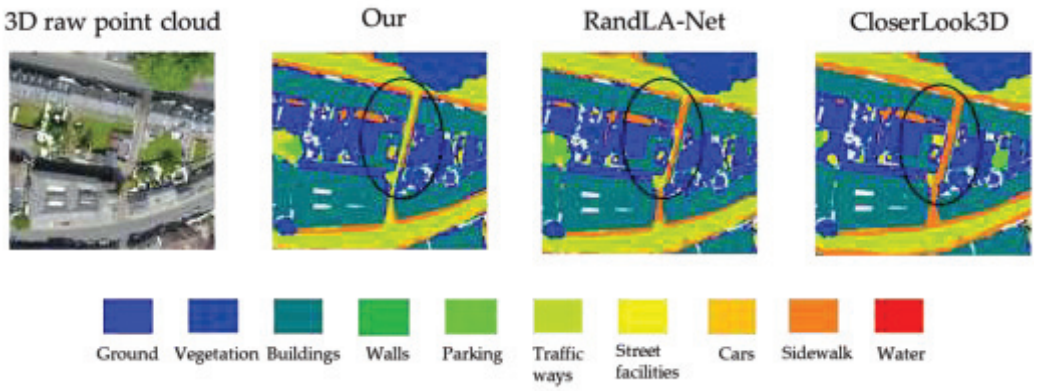
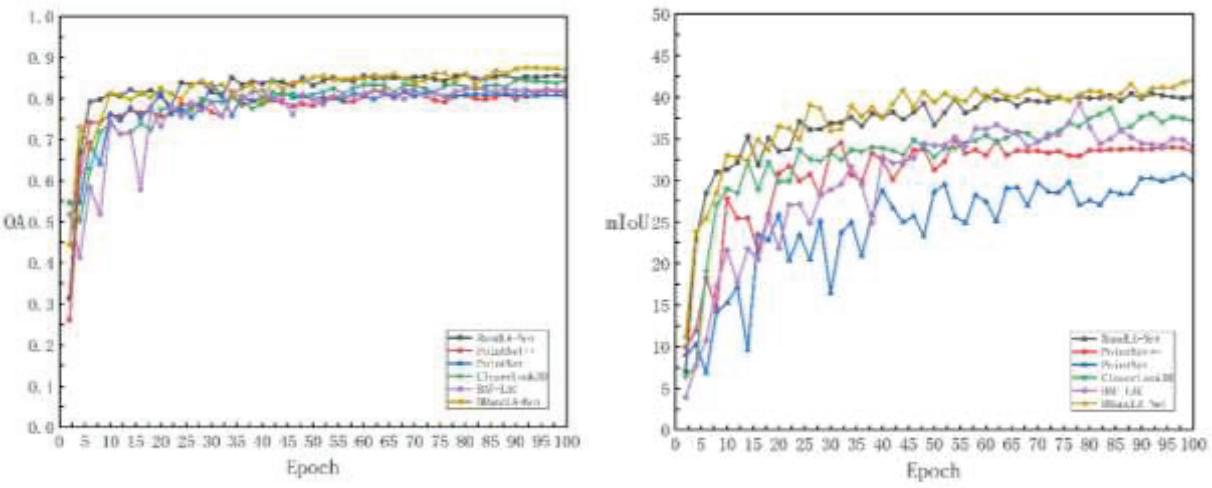


Figure 5. SensatUrban data set visualization results

3.2. Experimental results of classification of HPU dataset

To verify the generalization ability of the proposed DRandLA-Net algorithm, a comparative analysis was conducted with other point cloud segmentation models including PointNet++, PointNet, BAF-LAC, CloserLook3D, and RandLA-Net on the HPU point cloud dataset. As Figure 6(a) and Figure 6(b) depict the overall accuracy OA and the average cross-merge ratio mIoU versus the number of traversals Epoch. It can be observed from the figures that when the network model was trained for 85 epochs, the overall accuracy of DRandLA-Net tended to be stable and was better than the other five methods. When the number of epochs reached 90, the overall accuracy of DRandLA-Net reached its optimum and was improved by 5.26%, 3.02%, and 2.28% compared to the point cloud segmentation methods (BAF-LAC, CloserLook3D, Rand-LA-Net) suitable for large scenes. Compared to classical point cloud segmentation models PointNet++ and PointNet, the overall accuracy was improved by 5.98% and 6.85%, respectively. The mIoU also increased with the increase in training iterations and finally stabilized at around 42.54% Compared to BAF-LAC, CloserLook3D, and RandLA-Net, the mIoU was improved by 8.08%, 5.44%, and 2.94%, respectively. Com-pared to PointNet++ and PointNet, the mIoU was improved by 9.14% and 12.01%, respectively. These results demonstrate the superior generalization ability of the pro-posed DRandLA-Net algorithm in point cloud segmentation.



(a) (b)

Figure 6. (a) describes the relationship between the overall accuracy OA and the number of tra-versals Epoch; (b) describes the relationship between the average intersection ratio mIoU and the number of traversals Epoch

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202305_9(5).0007

Table 3 presents the quantitative results of the six methods on the 9-class semantic segmentation of the HPU dataset. The proposed DRandLA-Net algorithm achieved the best overall accuracy (OA) and mean intersection-over-union (mIoU) on the experimental dataset, with an overall accuracy of 87.42%, and the IoU accuracies of ground, vegetation, and building were 68.50%, 84.26%, and 89.73%, respectively. Compared to the PointNet and PointNet++ segmentation networks, DRandLA-Net utilizes a deep local feature aggregation module to extract relevant features between points, which can achieve better segmentation results. Compared to the point cloud segmentation methods (BAF-LAC, CloserLook3D, and RandLA-Net) suitable for large scenes, DRandLA-Net uses skip connections to fuse multi-scale features, and combines them with the deep local feature aggregation module, which makes the network better at capturing fine-grained details and coarse-grained semantic information of small target land classes, improving the semantic segmentation accuracy of small target land classes. The segmentation accuracies of walls, street facilities, cars, and sidewalks, four small target objects, were 17.4%, 7.51%, 44.48%, and 12.49%, respectively. Compared to the segmentation accuracies of BAF-LAC, CloserLook3D, and RandLA-Net, the seg-mentation accuracies of walls were improved by 5.27%, 4.52%, and 0.31%), respectively; the segmentation accuracies of street facilities were improved by 4.52%, 0.31%, and 1.27%, respectively; the segmentation accuracies of cars were improved by 3.75%, 5.72%, and 1.49%, respectively; and the segmentation accuracies of sidewalks were improved by 12.49%, 7.07%, and 6.35%, respectively. These results indicate that the proposed DRandLA-Net algorithm is feasible and effective for semantic segmentation of small target objects in point clouds.

Table 3. Evaluation of segmentation accuracy of different segmentation methods

Models	IoU(%)								
	Grounds	Vegetation	Buildings	Walls	Traffic ways	Street facilities	Cars	Sidewalks	Water
CloserLook3D	65.18	79.84	84.46	10.16	39.46	7.20	38.76	5.42	3.12
RandLA-Net	67.92	83.30	87.72	13.33	48.27	6.24	42.99	6.14	0.45
PointNet++	62.30	80.15	83.11	2.72	34.58	3.66	30.46	2.15	1.48
BAF-LAC	60.08	76.22	82.19	12.13	35.26	2.99	40.73	0.00	0.55
PointNet	58.40	76.24	76.36	0.00	34.48	0.00	29.31	0.00	0.00
Our	**68.50**	**84.26**	**89.73**	**17.40**	**53.48**	**7.51**	**44.48**	**12.49**	**5.06**

Table 4 presents the quantitative comparison results of our DRandLA-Net algorithm with other state-of-the-art point cloud segmentation models on the HPU dataset. The overall accuracy (OA) of our algorithm on the HPU dataset reached 87.42%, and the mean intersection-over-union (mIoU) reached 42.54% Compared to the PointNet and PointNet++ models which directly classify point clouds, DRandLA-Net better con-siders the relative feature relationships between points, thus improving the segmentation effect. The overall accuracy was improved by 6.85% and 5.98%, and the mIoU value was improved by 12.01% and 9.14%, respectively. Compared to the point cloud classification networks for large scenes such as BAF-LAC, CloserLook3D, and Rand-LA-Net, DRandLA-Net method expands the receptive field by constructing a deep lo-cal feature aggregation module, uses skip connections to connect between the encoder and decoder and inside the decoder to achieve multi-scale feature fusion, comprehensively captures fine-grained details and coarse-grained semantic information, enhances the correlation between points, and effectively improves the 3D scene semantic segmentation ability. The overall accuracy was improved by 5.26%, 3.02%, 2.28%, and the mIoU value was improved by 8.08%, 5.44%, 2.94%, respectively. The experimental results

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202305_9(5).0007

demonstrate that our DRandLA-Net algorithm is suitable for complex scenes and has better segmentation performance for small land objects with imbalanced samples.

Table 4. Segmentation results on HPU dataset

Models	MIoU(%)	OA(%)
CloserLook3D	37.10	84.40
RandLA-Net	39.60	85.14
PointNet++	33.40	81.44
BAF-LAC	34.46	82.16
PointNet	30.53	80.57
Our	**42.54**	**87.42**

Figure 7 displays the visual comparison results of semantic segmentation of the test dataset by our DRandLA-Net algorithm and five other state-of-the-art models. From the black oval-marked part in the figure, it can be seen that our method has better segmentation performance for roads, ground, cars, and water bodies. Compared with the segmentation results of BAF-LAC, CloserLook3D, and RandLA-Net algorithms, DRandLA-Net algorithm has more accurate segmentation performance for water bodies and can also extract target objects more completely for small land objects such as cars and narrow roads. For the PointNet and PointNet++ algorithms, it can be clearly seen that these two methods have misclassification when segmenting ground, water bodies, and roads, and cannot extract point clouds of small target objects completely. As DRandLA-Net combines deep local feature aggregation modules and multiscale fusion modules, the improved model has stronger feature extraction ability between points, thus more accurately identifying small structures on the top of buildings and other detailed information of other land cover types. The experimental results once again demonstrate that our DRandLA-Net algorithm is suitable for complex scenes and has better segmentation performance for small land objects with imbalanced samples.

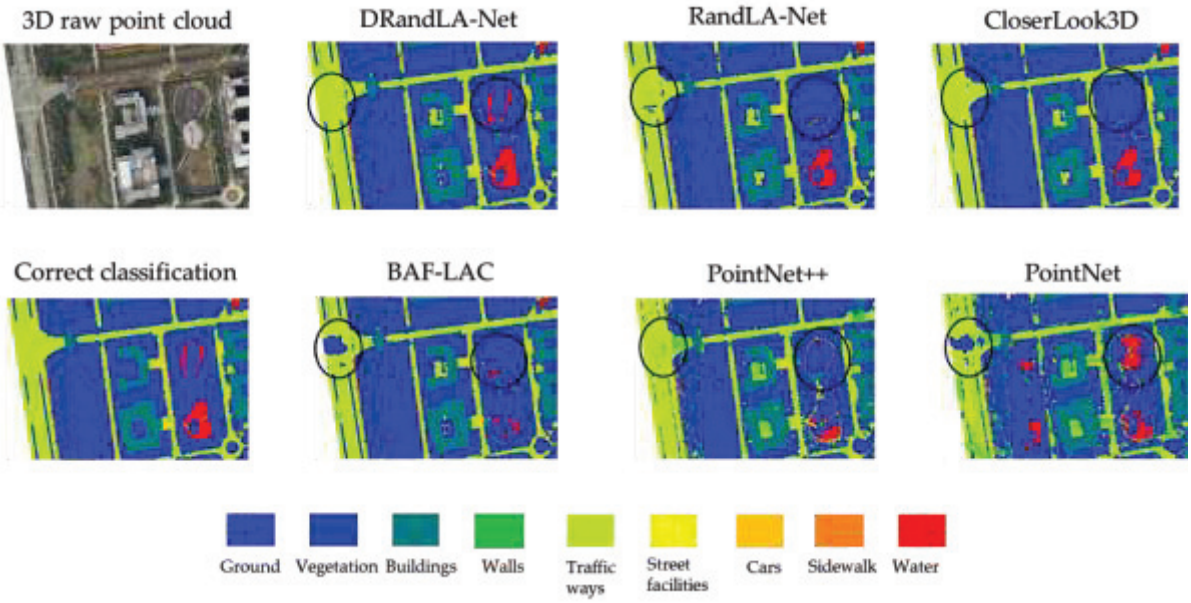


Figure 7. Semantic segmentation results for HPU dataset

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202305_9(5).0007

3.3. Ablation experiment

To evaluate the effectiveness of the deep local feature aggregation module (DLFA) proposed in this paper, we conducted segmentation experiments by incorporating the module into the BAF-LAC [24] neural network model, which is also designed for large-scale point cloud segmentation tasks. The results in Table 5 demonstrate that after embedding the DLFA module, the overall accuracy (OA) of the BAF-LAC network for segmentation increased by 1.32%, and the mean intersection-over-union (mIoU) increased by 0.82% Similarly, the overall accuracy of the RandLA-Net network for segmentation increased by 2.28%, and the mIoU increased by 2.94% after embedding the DLFA module. These results indicate that the DLFA module proposed in this paper is effective and feasible in improving the segmentation performance of existing point cloud segmentation models.

Table 5. Ablation experiments

Models	DLFA	MIoU (%)	OA (%)
BAF-LAC		82.16	34.46
BAF-LAC+DLFA	√	**83.48**	**35.28**
RandLA-Net		85.14	39.60
RandLA-Net+DLFA	√	**87.42**	**42.54**

4. CONCLUSIONS

In summary, this paper proposes an improved DRandLA-Net model for semantic segmentation of large-scale photogrammetric point cloud data. The model combines a deep local feature aggregation module and a multi-scale fusion module to comprehensively capture fine-grained details and coarse-grained semantic information, and im-prove the feature correlation between points. It has better segmentation performance for complex scenes and small land objects with imbalanced samples. Experimental results show that the proposed DRandLA-Net model achieved the best accuracy on the SensatUrban public point cloud dataset and the HPU dataset, with overall accuracy (OA) of 90.97% and 87.42%, and mean intersection-over-union (mIoU) of 53.84% and 42.54%, respectively. In addition, we verify the effectiveness of the deep local feature aggregation module and embed it into two neural network models, BAF-LAC and RandLA-Net, achieving further accuracy improvement. Therefore, the proposed DRandLA-Net model provides a more effective and feasible solution for semantic segmentation of large-scale point cloud data in complex scenes.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (U1304402).

REFERENCES

[1] Y. Wang, J. Wang, S. Chang, et al: Classification of street tree species using UAV tilt photogrammetry. Remote Sensing,Vol.13(2021) No.2, p. 216-234.

[2] Y. Xie, J. Tian, X.X. Zhu. Linking points with labels in 3D: A review of point cloud semantic segmentation. IEEE Geoscience and remote sensing magazine, Vol.8 (2020) No.4, p.38-59.

[3] X. Zou, M. Cheng, C. Wang, et al: Tree classification in complex forest point clouds based on deep learning. IEEE Geoscience and Remote Sensing Letters, Vol.14 (2017) No.12, p.2360-2364.

[4] J. Chen, Y. Chen, Z. Liu: Classification of Typical Tree Species in Laser Point Cloud Based on Deep Learning. Remote Sensing, Vol.13 (2021) No.23, p.4750-4771.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202305_9(5).0007

[5] A. Diab, R. Kashef, A. Shaker: Deep Learning for LiDAR Point Cloud Classification in Remote Sensing. Sensors, Vol.22 (2022) No.20, p.7868-7882.

[6] D. Ren, Z. Wu, J. Li, et al: Point attention network for point cloud semantic segmentation. Science China Information Sciences, Vol.65 (2022) No.192104, p.1-14.

[7] H. Su, S. Maji, E. Kalogerakis, et al: Multi-View convolutional neural networks for 3d shape recognition. IEEE Computer Society, Vol. 114 (2015) No.537, p.945-953.

[8] A. Boulch, J. Guerry, S.B. Le, et al: SnapNet: 3d point cloud semantic labeling with 2D deep segmentation networks. Computers & Graphics, Vol.71 (2018) No9, p.189-198.

[9] C.R. Qi, Y. Li, H. Su, et al: Pointnet++: deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems, Vol.30 (2017) No.8, p.5099-5188.

[10] C.R. Qi, H. Su, K. Mo, et al: Pointnet: deep learning on point sets for 3d classification and segmentation, IEEE Computer Society, Vol.16 (2017) p.77-85.

[11] K.G. Zhuang, M. Hao, J. Wang, et al: Linked dynamic graph CNN: learning on point cloud via linking hierarchical features. IEEE Trans, Vol.41 (2019) No.5 p.1939-1952.

[12] G. Te, W. Hu, A.M. Zheng, et al: Rgcnn: regularized graph CNN for point cloud segmentation, Association for Computing Machinery, (2018) p.746-754.

[13] C. Chen, L.Z. Fragonara, A. Tsourdos: Gapnet: graph attention based point neural network for exploiting local feature of point cloud. IEEE Access, Vol.8 (2020) No.6 p.148999-149012.

[14] Y.F. Xu, T.Q. Fan, M.Y. Xu, et al: Spidercnn: deep learning on point sets with parameterized convolutional filters, Proceedings of the Euro- pean Conference on Computer Vision, Vol.11212 (2018) No.2649, p.99-105.

[15] Q.Y. Hu, B. Yang, L.H. Xie, et al: Randla-net: efficient semantic segmentation of large-scale point clouds, Pro-ceedings of the IEEE, Vol.01112 (2020) No.20, p.11105-11114.

[16] Y. Chen, Y. Xiong, B. Zhang, et al: 3D point cloud semantic segmentation toward large-scale unstructured agricultural scene classification. Computers and Electronics in Agriculture, Vol.190 (2021) p.106445-106455.

[17] X. Han, Z. Dong and B. Yang: A point-based deep learning network for semantic segmentation of MLS point clouds. ISPRS Journal of Photogrammetry and Remote Sensing, Vol.175 (2021) No.4 p.199-214.

[18] J. Du: ResDLPS-Net: Joint residual-dense optimization for large-scale point cloud semantic segmentation. ISPRS Journal of Photogrammetry and Remote Sensing, Vol.182 (2021) p.37-51.

[19] J. Chen, Y. Zhao, C. Meng, et al: Multi-Feature Aggregation for Semantic Segmentation of an Urban Scene Point Cloud. Remote Sensing, Vol.14 (2022) No.20 p.5134-5151.

[20] J. Long, E. Shelhamer, T. Darrell: Fully convolutional networks for semantic segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.39 (2015) No.4 p.640-651.

[21] O. Ronneberger, P. Fischer, T. Brox: U-net: Convolutional networks for biomedical image segmentation. CoRR, Vol. 9351 (2015) p.234-241.

[22] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, et al: Unet++: A nested u-net architecture for medical image segmentation, Multimodal Brain Image Analysis, Vol. 11045 (2018) pp.3-11.

[23] Q.Y. Hu, B. Yang, S. Khalid, et al: SensatUrban: learning semantics from urban-scale photogrammetric point clouds. Internati- onal Journal of Computer Vision, Vol.15 (2022) No.8 p.1-28.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202305_9(5).0007

[24] H. Shuai, X. Xu, Q. Liu: Backward attentive fusing network with local aggregation classifier for 3d point cloud semantic seg-mentation. IEEE Transactions on Image Processing, Vol.30 (2021) No.6 p.4973-4984.